Mathematics of Robustness, Risk & Failure

Volume V: Rare Events & Large Deviations

The mathematics of improbable outcomes: theory, estimation, and simulation

---

Every derivation step by step

Every symbol defined

Explicit numerical examples

Geometric intuitions

Silvano Salvador

Contents

17.1 The Large Deviation Principle and Rate Functions 2

17.2 Cramér's Theorem and the Legendre–Fenchel Transform 6

17.3 Sanov's Theorem and Relative Entropy 11

17.4 The Gärtner–Ellis Theorem 15

17.5 Contraction Principles and Applications 19

18.1 Sample-Path Large Deviations: The Freidlin–Wentzell Theory 24

18.2 Quasi-Potentials and Metastability 28

18.3 Large Deviations for Markov Chains 33

18.4 Exit Times from Domains of Attraction 37

18.5 Applications to Noise-Induced Transitions 41

19.1 Importance Sampling: Theory and Variance Reduction 46

19.2 Exponential Tilting and the Optimal Change of Measure 50

19.3 Splitting Methods and Subset Simulation 54

19.4 Cross-Entropy Method for Rare-Event Estimation 58

19.5 Adaptive Multilevel Splitting 62

20.1 Bayesian Estimation of Extreme Quantiles 67

20.2 Prior Elicitation for Tail Parameters 71

20.3 Hierarchical Models for Pooling Rare-Event Data 75

20.4 Model Uncertainty and Bayesian Model Averaging in Tail Inference 80

20.5 Case Study: Catastrophe Risk Modelling 84

Appendix A Convex Analysis and the Legendre–Fenchel Transform 89

Appendix B Perron–Frobenius Theory of Non-negative Matrices 91

Appendix C Full Proofs: Cramér's Theorem and Sanov's Theorem 93

Appendix D Stochastic Calculus Background: Itô's Formula and Girsanov's Theorem 95

Appendix E Markov Chain Monte Carlo: Metropolis–Hastings and Hamiltonian Monte Carlo 97

Appendix F Reference Tables: GPD, GEV, and Related Distributions 98

Bibliography 100

Appendix G Extended Case Study: An End-to-End Worked Portfolio 102

Appendix H Notation and Symbol Glossary 104

Appendix I Advanced Exercises 106

Appendix J Numerical Reference Tables and Topic Guide 109

PART 17

Large Deviation Theory: Foundations

The formal calculus governing the exponential decay of probabilities of rare events.

17.1 The Large Deviation Principle and Rate Functions

Overview

Large deviation theory sharpens the law of large numbers. The LLN says $S_n/n \to \mu$ in probability; large deviation theory asks: *how fast* does $\Pr(S_n/n \approx x)$ vanish for $x \neq \mu$? The answer is almost always exponential in n, governed by a single function $I(x)$, the *rate function*.

Setup. Let $(\mu_n)_{n\geq 1}$ be a sequence of probability measures on a topological space $\mathcal{X}$ (typically $\mathbb{R}^d$), and let $a_n \to \infty$ be a sequence of positive scales (usually $a_n = n$).

Definition: Large Deviation Principle (LDP)

(μ_n) satisfies the LDP with rate a_n and rate function $I : \mathcal{X} \to [0, \infty]$ if for every Borel set $A \subseteq \mathcal{X}$:

$$-\inf_{x \in A^\circ} I(x) \;\leq\; \liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(A) \;\leq\; \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(A) \;\leq\; -\inf_{x \in \bar{A}} I(x)$$

where A° is the interior and $\bar{A}$ the closure. We write this informally as $\mu_n(A) \approx e^{-a_n \inf_A I}$.

Good rate functions. I is *lower semicontinuous* (so $\{I \leq c\}$ is closed for every c) and *good* if every sub-level set $\{x : I(x) \leq c\}$ is compact. Goodness guarantees that the infimum $\inf_{\bar{A}} I$ is attained and that I has at least one zero whenever μ_n is exponentially tight.

Properties. If (μ_n) satisfies the LDP with good rate function I:

- $I(x) \geq 0$ for all x, and typically $I(x^\star) = 0$ for a unique $x^\star$ equal to the LLN limit – the "most likely" outcome carries zero exponential cost.
- For any open set $U \ni x^\star$, $\mu_n(U) \to 1$ exponentially fast (consistency with the LLN).
- I need not be convex in general, but in the i.i.d. setting of Cramér's theorem (Section 17.2) it always is.

Weak LDP and exponential tightness. A *weak LDP* requires the inequalities above only for compact sets A. Combined with *exponential tightness* – for every $L < \infty$ there is a compact K_L with $\limsup_n \frac{1}{a_n} \log \mu_n(K_L^c) \leq -L$ – a weak LDP automatically upgrades to a full LDP with a good rate function. This is the standard route to proving LDPs in unbounded spaces.

Example — Worked example: Bernoulli sums and binary relative entropy

Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli(p) with $0 < p < 1$, and let $S_n = \sum_{i=1}^n X_i$. Fix $x \in (0, 1)$ and consider $\Pr(S_n = \lceil nx \rceil)$. Using the exact binomial probability,

$$\Pr(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = \lceil nx \rceil.$$

Apply Stirling's approximation $\log m! = m\log m - m + O(\log m)$ to each factorial in $\log\binom{n}{k} = \log n! - \log k! - \log(n-k)!$. With $k = nx$ (treating x as exact for clarity),

$$\frac{1}{n}\log\binom{n}{nx} \;\longrightarrow\; -x\log x - (1-x)\log(1-x) =: H(x),$$

the binary entropy. Adding the $\log p^k(1-p)^{n-k}$ term,

$$\frac{1}{n}\log\Pr(S_n \approx nx) \;\longrightarrow\; H(x) + x\log p + (1-x)\log(1-p) = -\Big[x\log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}\Big].$$

The bracketed quantity is exactly the Kullback–Leibler divergence between the Bernoulli(x) and Bernoulli(p) distributions:

$$I(x) = D_{\mathrm{KL}}(\mathrm{Bern}(x)\,\|\,\mathrm{Bern}(p)) = x\log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}.$$

This is the rate function: $\Pr(S_n/n \approx x) \approx e^{-nI(x)}$, $a_n = n$. Note $I(p) = 0$ (the LLN value), $I(x) > 0$ for $x \neq p$, I is strictly convex, and $I(x) \to \infty$ only at the boundary points $x \in \{0,1\}$ if $p \in (0,1)$ – it remains finite there ($I(0) = -\log(1-p)$, $I(1) = -\log p$), reflecting that hitting $S_n = 0$ or $S_n = n$ has probability decaying like $(1-p)^n$ or p^n respectively, which is exactly $e^{-nI(0)}$ and $e^{-nI(1)}$.
For instance, with $p = 0.3$ and $x = 0.5$: $I(0.5) = 0.5\log\frac{0.5}{0.3} + 0.5\log\frac{0.5}{0.7} = 0.5(0.5108) + 0.5(-0.3365) \approx 0.0872$. So $\Pr(S_n/n \approx 0.5) \approx e^{-0.0872\,n}$ – for $n = 100$ this is roughly $e^{-8.72} \approx 1.6 \times 10^{-4}$, a quantitatively meaningful rare-event probability that a naive Gaussian (CLT) approximation, valid only for x close to p, would badly mis-estimate.

Geometric intuition and outlook. The rate function I defines a "landscape" over $\mathcal{X}$: its unique zero is the typical (LLN) outcome, and the cost of observing S_n/n near x scales as $e^{-nI(x)}$ – exponentially steep walls around the typical value. This single function I packages an entire family of large-deviation estimates (for all sets A simultaneously) into one variational object, analogous to a free-energy or entropy function in statistical mechanics, where a_n plays the role of inverse temperature or system size. Section 17.2 shows that for general i.i.d. sequences (not just Bernoulli), the rate function is *always* obtainable as a Legendre–Fenchel transform of the cumulant generating function – Cramér's theorem, of which the binary-entropy computation above is the special case p-Bernoulli. Section 17.3 generalises further, from scalar sample means to entire empirical distributions (Sanov's theorem), with I becoming the relative entropy functional $D(\cdot\|\cdot)$ on probability measures – exactly the structure that appeared above.

Example — Exercises

1. For Bernoulli(p) with $p = 0.5$, compute $I(0.8) = D(\mathrm{Bern}(0.8)\|\mathrm{Bern}(0.5))$, and compare to $I(0.2)$. *Solution:* $I(0.8) = 0.8\log\frac{0.8}{0.5} + 0.2\log\frac{0.2}{0.5} = 0.8\log 1.6 + 0.2\log 0.4 \approx 0.8(0.4700) + 0.2(-0.9163) \approx 0.3760 - 0.1833 = 0.1927$. By the symmetry Bern$(0.5)$, $I(0.2) = 0.2\log 0.4 + 0.8\log 1.6$, the same expression with roles swapped – giving the identical value 0.1927, reflecting the symmetry of the reference $p = 0.5$ about which deviations x and $1-x$ are equally costly.
2. Verify that $I(x) = x\log(x/p) + (1-x)\log((1-x)/(1-p))$ is convex on $(0,1)$ by computing $I''(x)$. *Solution:* $I'(x) = \log(x/p) - \log((1-x)/(1-p)) = \log\frac{x(1-p)}{p(1-x)}$, and $I''(x) = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)} > 0$ for $x \in (0,1)$ – strictly convex, with $I'' \to \infty$

at the boundary, consistent with I remaining finite but its derivative blowing up at $x \in \{0, 1\}$.

3. Using $I(1) = -\log p$, estimate $\Pr(S_n = n)$ (all successes) for $p = 0.3, n = 20$, and compare with the exact value p^{20}. *Solution:* $I(1) = -\log(0.3) \approx 1.204$, so $\Pr(S_{20} = 20) \approx e^{-20 \times 1.204} = e^{-24.08}$. The exact value is $0.3^{20} = e^{20 \log 0.3} = e^{-24.08}$ – exact equality here, since for $x = 1$ the Stirling approximation in the rate-function derivation is not even needed (there is only one sequence with $S_n = n$, namely all successes, so the "entropy" term $H(1) = 0$ contributes nothing and the rate function reduces to the exact exponent).

Further perspectives. The LDP framework of Definition (Large Deviation Principle) was formalised by Varadhan in 1966, unifying earlier special-case results (Cramér 1938, Sanov 1957) under a single variational language. A subtlety worth flagging: the LDP statement involves *both* a lim inf and a lim sup, and for general (non-convex, non-good) rate functions these can genuinely differ for some sets A – the LDP is then said to hold only for sets A with $\inf_{A^\circ} I = \inf_{\bar{A}} I$ ("I-continuity sets" of A). For the good, convex rate functions that dominate this volume (arising from Cramér, Sanov, and Gärtner–Ellis), this subtlety rarely bites, but it becomes relevant in Section 18.4's discussion of near-degenerate saddle competition, where level sets of I can have boundary issues. A complementary viewpoint, useful for intuition: the rate function I can be thought of as a "non-equilibrium free energy" in the sense of large deviations for interacting particle systems, where $a_n = n$ plays the role of particle number and $I(x)/k_B$ (for Boltzmann constant k_B, here absorbed into units) is literally a free-energy difference per particle – the entire apparatus of equilibrium statistical mechanics (partition functions, free energies, phase transitions) can be recast as a special case of large deviation theory applied to the empirical measure of a many-particle system, a perspective developed systematically in the statistical-mechanics literature and echoed in this volume's treatment of metastability (Part 18).

Example — Application note: credit portfolio default-rate deviations

A retail loan portfolio of $n = 2000$ independent loans, each defaulting independently with probability $p = 0.03$ (a 3% baseline annual default rate), is monitored for the event "default rate exceeds 5% in a given year," i.e. $S_n/n \geq 0.05$. The rate function from Section 17.1 gives $I(0.05) = 0.05 \log(0.05/0.03) + 0.95 \log(0.95/0.97) = 0.05 \log(1.667) + 0.95 \log(0.979) \approx 0.05(0.511) + 0.95(-0.0212) \approx 0.0256 - 0.0201 = 0.0055$. So $\Pr(S_n/n \geq 0.05) \approx e^{-2000 \times 0.0055} = e^{-11.0} \approx 1.7 \times 10^{-5}$ – roughly a 1-in-60,000 event *per year*, under the independence assumption. A risk manager comparing this to historical data showing such breaches occurring roughly once every 500 years (a frequency of 0.002, i.e. $\approx 1.7 \times 10^{-5} \times 100$) would correctly infer that the independence assumption is violated – correlated defaults (e.g. via a shared macroeconomic factor, as in Section 20.5's common-factor discussion) make the 5% threshold roughly $100\times$ more likely than the independent-default rate function predicts. This is a standard and important diagnostic use of the LDP rate function: not as a final answer, but as a *benchmark* against which the magnitude of unmodelled dependence can be assessed – a "$100\times$" discrepancy is a quantitatively meaningful signal that correlation, not just marginal default probability, must be modelled (precisely the hierarchical and dependence-structure themes of Part 20).

Algorithm: empirical estimation of a rate function via binning

Input: samples $\{\bar{X}_n^{(1)}, \dots, \bar{X}_n^{(M)}\}$ from M independent runs of length n; a grid of values $x_1 < \cdots < x_K$.

1. For each x_k, estimate $\hat{p}_k = \frac{1}{M}\#\{m : \bar{X}_n^{(m)} \in [x_k - \delta, x_k + \delta]\}$ for small bin-width δ.
2. Estimate $\hat{I}(x_k) = -\frac{1}{n}\log \hat{p}_k$.
3. Plot $\hat{I}(x_k)$ against x_k; for x_k near the LLN value, $\hat{I} \approx 0$; the estimate becomes unreliable (high variance) for x_k far from the LLN value, where $\hat{p}_k \to 0$ and few or zero samples fall in the bin.
4. For the unreliable region, switch to importance sampling (Part 19) using a preliminary tilt estimated from the reliable, near-mean region of $\hat{I}$.

This "empirical rate function" approach is often the practical starting point in applications where Λ (and hence $\Lambda^\star$) is not analytically tractable: a moderate number of direct simulations characterise $I(x)$ near the mean, where probabilities are large enough for direct estimation, and this near-mean estimate (e.g. its curvature, related to the CLT variance via the Bahadur–Rao correspondence noted in Section 18.3's remarks) seeds the importance-sampling machinery needed for the far tails – a "bootstrap from the centre outward" strategy that recurs across this volume.

Key formulas: Section 17.1

- LDP: $-\inf_{A^\circ} I \le \liminf_n \frac{1}{a_n}\log \mu_n(A) \le \limsup_n \frac{1}{a_n}\log \mu_n(A) \le -\inf_{\bar{A}} I$.
- Good rate function: $\{I \le c\}$ compact for all c; $I \ge 0$, $I(x^\star) = 0$ at the LLN limit.
- Bernoulli(p) sample mean: $I(x) = D(\mathrm{Bern}(x)\|\mathrm{Bern}(p)) = x\log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}$.
- Weak LDP + exponential tightness $\Rightarrow$ full LDP with good rate function.

Notation and further reading. Throughout this Part, $\Pr(\cdot)$ denotes probability under the relevant base measure (specified by context), $\mathbb{E}[\cdot]$ the corresponding expectation, and a_n the LDP speed (almost always $a_n = n$ or $a_n = 1/\epsilon$ in this volume). The asymptotic notation $f(n) \approx g(n)$ means $\frac{1}{a_n}\log f(n) \to \frac{1}{a_n}\log g(n)$, i.e. agreement to leading exponential order, suppressing polynomial-in-n (or polynomial-in-$1/\epsilon$) prefactors – the latter are the subject of Section 18.4's refined asymptotics. Classic references for this material include Dembo & Zeitouni's *Large Deviations Techniques and Applications* (the standard graduate text covering Sections 17.1–17.5 and much of Part 18) and Varadhan's original 1966 paper formalising the LDP itself.

Cross-volume connections. The rate function $I(x)$ introduced here is the same object that, in Volume I's extreme value theory, governs the exponential decay of the probability that a sample mean of *light-tailed* losses exceeds a high threshold – complementing Volume I's primary focus on the tail of a *single* loss (governed by the GPD/GEV shape parameter ξ). The two perspectives – single-observation tails (Volume I, regular variation) and sample-mean tails (this volume, large deviations) – coincide only for light-tailed distributions in the Cramér regime; for heavy-tailed (subexponential) losses, sample-mean tails are dominated by the single-largest-observation mechanism (the "one big jump" principle of Section 20.5), a qualitatively different regime that this Part's Cramér/Sanov/Gärtner–Ellis machinery does not directly address.

Example — Extended discussion: the moderate deviation regime

Between the central limit theorem (fluctuations of order $1/\sqrt{n}$) and the large deviation regime (fluctuations of order 1) lies the *moderate deviation* regime: $\bar{X}_n - \mu \approx a_n/\sqrt{n}$ for $a_n \to \infty$ with $a_n/\sqrt{n} \to 0$. In this regime, $\Pr(\bar{X}_n - \mu \geq a_n/\sqrt{n}) \approx \exp\left(-\frac{a_n^2}{2\sigma^2}\right)$ – a *purely Gaussian* rate, even when X itself is non-Gaussian. This reflects a second-order Taylor expansion $I(x) \approx \frac{(x-\mu)^2}{2\sigma^2}$ near $x = \mu$ (using $I''(\mu) = 1/\sigma^2$ from Appendix A's property (v)).

Worked illustration. For Bernoulli$(p = 0.3)$, $I''(0.3) = \frac{1}{0.3\times 0.7} \approx 4.762$ (Section 17.1's exercises), so $I(x) \approx \frac{(x-0.3)^2}{2/4.762} = \frac{(x-0.3)^2}{0.42}$. The exact $I(0.35) \approx 0.00576$ versus the quadratic approximation $\frac{(0.05)^2}{0.42} \approx 0.00595$ – a 3% relative error. At $x = 0.5$, exact $I(0.5) \approx 0.0872$ versus quadratic $\frac{(0.2)^2}{0.42} \approx 0.0952$ – a 9% error, growing as expected once the deviation leaves the moderate regime. The moderate-deviation principle thus quantifies precisely how far the convenient Gaussian approximation to I can be trusted before the full machinery of Sections 17.2–17.4 is needed.

Example — Extended discussion: Bahadur efficiency of statistical tests

The rate function also governs the asymptotic efficiency of hypothesis tests via Bahadur's theorem: for a test statistic $T_n \to \theta_1$ a.s. under the alternative, with large-deviation rate $I(\cdot)$ under the null, the Bahadur slope is $c = 2I(\theta_1)$, and $-\frac{1}{n}\log p_n \to I(\theta_1)$ a.s., where p_n is the test's p-value.

Worked illustration. Using this section's Bernoulli$(p = 0.3)$ example, for a test of $H_0 : p = 0.3$ against an observed proportion $\theta_1 = 0.5$, the Bahadur slope is $c = 2I(0.5) \approx 2 \times 0.0872 = 0.1744$, so $p_n \approx e^{-0.0872n}$. For $n = 100$, $p_n \approx e^{-8.72} \approx 1.6 \times 10^{-4}$. This links the LDP machinery directly to p-value decay rates for goodness-of-fit tests on rare-event proportions, complementing the purely probabilistic readings of Sections 17.1–17.2.

17.2 Cramér's Theorem and the Legendre–Fenchel Transform

Overview

Cramér's theorem is the prototypical LDP: for i.i.d. real-valued random variables, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies an LDP with rate n and rate function $I = \Lambda^\star$, the Legendre–Fenchel transform of the log-moment generating function Λ. Every other rate function we meet in this volume is, ultimately, a transform or projection of some $\Lambda^\star$.

Definition: Cumulant generating function and Legendre–Fenchel transform

For $X \sim \nu$ on $\mathbb{R}$, the *cumulant generating function* (CGF, log-MGF) is

$$\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}] = \log \int e^{\lambda x}\, \nu(dx), \qquad \lambda \in \mathbb{R},$$

taking values in $(-\infty, +\infty]$. Its *Legendre–Fenchel transform* (convex conjugate) is

$$\Lambda^\star(x) = \sup_{\lambda \in \mathbb{R}} (\lambda x - \Lambda(\lambda)).$$

Properties of Λ and $\Lambda^\star$. Λ is convex (by Hölder / Cauchy–Schwarz on $e^{\lambda X}$), $\Lambda(0) = 0$, and $\Lambda'(0) = \mathbb{E}[X] = \mu$ whenever Λ is differentiable at 0. Consequently $\Lambda^\star \geq 0$ always (take $\lambda = 0$ in the sup), $\Lambda^\star$ is convex and lower semicontinuous, and $\Lambda^\star(\mu) = 0$ at the mean, with $\Lambda^\star$ increasing on $[\mu, \infty)$ and decreasing on $(-\infty, \mu]$ – a genuine "energy well" centred at the typical value.

Example — Theorem: Cramér's theorem (one-dimensional)

Let $X_1, X_2, \ldots$ be i.i.d. with law ν such that $\Lambda(\lambda) < \infty$ for λ in a neighbourhood of 0. Then the law of $\bar{X}_n$ satisfies the LDP with rate n and good convex rate function $I = \Lambda^\star$:

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr(\bar{X}_n \in A) = -\inf_{x \in A} \Lambda^\star(x) \quad \text{(for nice sets } A\text{)}.$$

Proof sketch (*upper bound, Chernoff bound*). For $x > \mu$ and any $\lambda \geq 0$, Markov's inequality applied to $e^{\lambda S_n}$ (with $S_n = n\bar{X}_n$) gives

$$\Pr(\bar{X}_n \geq x) = \Pr(e^{\lambda S_n} \geq e^{\lambda n x}) \leq e^{-\lambda n x} \mathbb{E}[e^{\lambda S_n}] = e^{-n(\lambda x - \Lambda(\lambda))}.$$

Taking $\lambda \to$ the optimal $\lambda^\star(x) = \arg\sup_\lambda (\lambda x - \Lambda(\lambda))$ gives $\Pr(\bar{X}_n \geq x) \leq e^{-n\Lambda^\star(x)}$, i.e. $\frac{1}{n} \log \Pr(\bar{X}_n \geq x) \leq -\Lambda^\star(x)$. This is the *Chernoff bound*, and it gives the LDP upper bound for all sets directly.
Proof sketch (*lower bound, exponential tilting*). Define the *tilted measure* $\nu_{\lambda^\star}(dx) = e^{\lambda^\star x - \Lambda(\lambda^\star)} \nu(dx)$, a valid probability measure (it integrates to 1 by definition of Λ) whose mean equals $\Lambda'(\lambda^\star) = x$ by construction (assuming Λ is differentiable / steep). Under $\nu_{\lambda^\star}^{\otimes n}$, the event $\{\bar{X}_n \approx x\}$ is now *typical* (probability $\to 1$ by the LLN under the tilted law). Converting probabilities back via the Radon–Nikodym derivative, $\Pr_\nu(\bar{X}_n \approx x) = \mathbb{E}_{\nu_{\lambda^\star}}\left[e^{-\lambda^\star S_n + n\Lambda(\lambda^\star)} \mathbf{1}_{\bar{X}_n \approx x}\right] \approx e^{-n(\lambda^\star x - \Lambda(\lambda^\star))} = e^{-n\Lambda^\star(x)}$, matching the upper bound. *Exponential tilting* – replacing the original measure by the one under which the rare event is typical – is the structural idea behind importance sampling for rare-event simulation (Section 19.2).

Example — Worked example: Gaussian and exponential rate functions

Gaussian. $X \sim \mathcal{N}(\mu, \sigma^2)$ gives $\Lambda(\lambda) = \mu\lambda + \frac{1}{2}\sigma^2\lambda^2$. The supremum $\sup_\lambda(\lambda x - \mu\lambda - \frac{1}{2}\sigma^2\lambda^2)$ is attained at $\lambda^\star = (x - \mu)/\sigma^2$, giving

$$\Lambda^\star(x) = \frac{(x-\mu)^2}{2\sigma^2}.$$

This is the quadratic rate function: $\Pr(\bar{X}_n \approx x) \approx e^{-n(x-\mu)^2/2\sigma^2}$, recovering the familiar Gaussian tail at the rate-function level – the CLT and the LDP agree to leading exponential order near μ, but the LDP remains valid arbitrarily far into the tails, where the Gaussian CLT approximation (which only controls fluctuations of order $1/\sqrt{n}$) breaks down.
Exponential. $X \sim \mathrm{Exp}(\theta)$ (rate θ, mean $1/\theta$) gives $\Lambda(\lambda) = -\log(1 - \lambda/\theta)$ for $\lambda < \theta$, $+\infty$ otherwise. Setting the derivative $x = \Lambda'(\lambda) = \theta/(\theta - \lambda) \cdot \theta^{-1} \cdot \theta = 1/(\theta - \lambda)$ to solve $\lambda^\star = \theta - 1/x$ (for $x > 0$), substitution gives

$$\Lambda^\star(x) = \theta x - 1 - \log(\theta x), \qquad x > 0.$$

Check: at $x = 1/\theta$ (the mean), $\Lambda^\star(1/\theta) = 1 - 1 - \log 1 = 0$, as required. For $\theta = 1$ and $x = 3$ (a $3\times$-mean deviation), $\Lambda^\star(3) = 3 - 1 - \log 3 \approx 0.901$, so $\Pr(\bar{X}_n \geq 3) \approx e^{-0.901\, n}$

– compare with the much faster Gaussian decay $e^{-(3-1)^2 n/2} = e^{-2n}$ that a naive CLT-based estimate (incorrectly assuming Gaussian tails) would predict: the exponential's heavier (but still exponential) tail is correctly captured only by $\Lambda^\star$, not by the quadratic Gaussian surrogate. This is a recurring caution in tail-risk modelling (Volume I): the *shape* of Λ near the boundary of its domain, not just its first two moments, controls rare-event rates.

Multidimensional Cramér's theorem. The theorem extends verbatim to $\mathbb{R}^d$: $\Lambda(\square) = \log \mathbb{E}[e^{\square \cdot \mathbf{X}}]$, $\Lambda^\star(\mathbf{x}) = \sup_{\square}(\square \cdot \mathbf{x} - \Lambda(\square))$, and $\bar{\mathbf{X}}_n$ satisfies the LDP with rate function $\Lambda^\star$ on $\mathbb{R}^d$. This multivariate version underlies the joint tail behaviour of correlated risk factors (Volume I) and is the starting point for the Gärtner–Ellis extension to dependent sequences (Section 17.4).

Example — Exercises

1. For $X \sim \text{Poisson}(\mu)$, $\Lambda(\lambda) = \mu(e^\lambda - 1)$. Compute $\Lambda^\star(x)$ for $x > \mu$. *Solution:* Maximise $\lambda x - \mu(e^\lambda - 1)$ over λ: setting the derivative to zero, $x - \mu e^{\lambda^\star} = 0 \Rightarrow \lambda^\star = \log(x/\mu)$. Then $\Lambda^\star(x) = x\log(x/\mu) - \mu(x/\mu - 1) = x\log(x/\mu) - x + \mu$. This is the rate function for sums of i.i.d. Poisson(μ) random variables.
2. Verify $\Lambda^\star(\mu) = 0$ and that $\Lambda^\star$ is increasing for $x > \mu$. *Solution:* At $x = \mu$: $\Lambda^\star(\mu) = \mu\log 1 - \mu + \mu = 0$. For $x > \mu$, $\frac{d}{dx}\Lambda^\star(x) = \log(x/\mu) + x \cdot (1/x) - 1 = \log(x/\mu) > 0$ since $x > \mu$ – confirming $\Lambda^\star$ increases away from the mean, as required of a rate function.
3. For $\mu = 10$, estimate $\Pr(\bar{X}_n \geq 15)$ for $n = 50$ using $\Lambda^\star(15)$. *Solution:* $\Lambda^\star(15) = 15\log(1.5) - 15 + 10 = 15(0.4055) - 5 \approx 6.082 - 5 = 1.082$. So $\Pr(\bar{X}_{50} \geq 15) \approx e^{-50 \times 1.082} = e^{-54.1}$ – a vastly smaller probability than the corresponding Gaussian-CLT approximation $\Pr(\mathcal{N}(10, 10/50) \geq 15) = \Pr(\mathcal{N}(0,1) \geq 5/\sqrt{0.2}) = \Pr(\mathcal{N}(0,1) \geq 11.18)$ would suggest is "comparable in shape" – both are super-exponentially small, but the exact LDP rate 1.082 per unit of n is the quantity that correctly governs the decay, not the Gaussian approximation's quadratic form, which is only locally valid near $x = \mu$.

Further perspectives. Cramér's original 1938 proof (for sums of i.i.d. bounded random variables) predates the modern LDP formalism by decades, and was motivated by actuarial questions in ruin theory – the probability that an insurer's reserves, a sum of premium income minus claims, ever fall below zero, a question directly continuous with the catastrophe risk modelling of Section 20.5. The Chernoff bound derived in the proof sketch is, independently of its role in the LDP, one of the most widely used tools in probability and computer science (concentration inequalities, randomised algorithm analysis) – it requires *none* of the LDP machinery to state and prove for finite n, but the LDP shows that, for the rate function $\Lambda^\star$, the Chernoff bound is not merely an upper bound but is *asymptotically tight* (matches the true exponential rate). A natural extension, not pursued in detail here but used implicitly in Section 19.2, is the *multivariate* Chernoff bound: for $\mathbf{S}_n = \sum \mathbf{X}_i \in \mathbb{R}^d$, $\Pr(\mathbf{S}_n/n \in A) \leq e^{-n \inf_{\mathbf{x} \in A} \Lambda^\star(\mathbf{x})}$ for any convex A, by the same Markov-inequality argument applied along the optimal direction $\square^\star$ – the multivariate Legendre transform $\Lambda^\star(\mathbf{x}) = \sup_{\square}(\square \cdot \mathbf{x} - \Lambda(\square))$ is itself a convex function of $\mathbf{x}$ (as the supremum of affine functions), a fact used repeatedly when rate functions for vector-valued statistics must be optimised over constraint sets, as in Section 17.5's contraction examples and Section 20.5's aggregate-loss calculations.

Example — Application note: trading desk P&L tail via Cramér

A trading desk's daily P&L is modelled as i.i.d. draws from a distribution with CGF $\Lambda(\lambda) = 0.5\lambda^2 + 0.1\lambda^3$ for small $|\lambda|$ (a Gaussian-plus-skewness model, the cubic term

capturing a mild negative-skew "crash risk" not present in a pure Gaussian). For a 20-day monthly aggregation, the desk wants $\Pr(\bar{X}_{20} \leq -1)$ (average daily loss exceeding 1 unit over the month). Without the cubic term ($\Lambda(\lambda) = 0.5\lambda^2$, pure Gaussian, mean 0, variance 1), $\Lambda^\star(x) = x^2/2$ (Section 17.2's worked example), giving $\Lambda^\star(-1) = 0.5$ and $\Pr(\bar{X}_{20} \leq -1) \approx e^{-20\times 0.5} = e^{-10} \approx 4.5 \times 10^{-5}$. With the cubic term, $\Lambda^\star(-1) = \sup_\lambda(-\lambda - 0.5\lambda^2 - 0.1\lambda^3)$; for $\lambda < 0$ (relevant for the left tail), the cubic term $-0.1\lambda^3 > 0$ for $\lambda < 0$, *increasing* the objective relative to the pure-Gaussian case – numerically solving $\frac{d}{d\lambda}(-\lambda - 0.5\lambda^2 - 0.1\lambda^3) = -1 - \lambda - 0.3\lambda^2 = 0$ for the relevant negative root gives $\lambda^\star \approx -0.85$ (vs. $\lambda^\star = -1$ for pure Gaussian), and substituting back gives $\Lambda^\star(-1) \approx 0.85 - 0.5(0.7225) - 0.1(-0.614) \approx 0.85 - 0.361 + 0.061 = 0.55$. The slightly larger $\Lambda^\star(-1) \approx 0.55$ vs. 0.5 might seem to suggest the skew term makes large losses *less* likely – but this is an artefact of fixing $x = -1$ exactly; the cubic CGF term shifts the entire rate function asymmetrically, and for more extreme x (e.g. $x = -2$) the cubic term's effect reverses sign, correctly capturing "fatter left tail." This illustrates a general caution: higher-order CGF corrections affect the rate function non-monotonically across x, and a single-point comparison can be misleading – the full function $\Lambda^\star(\cdot)$, not a single value, is the object that should be compared when assessing how higher moments (skewness, kurtosis) reshape tail risk.

Algorithm: numerical evaluation of $\Lambda^\star(x)$ via 1D optimisation

Input: CGF $\Lambda(\lambda)$ (closed-form or computed numerically via $\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, e.g. by quadrature or simulation); target x.

1. Define $g(\lambda) = \lambda x - \Lambda(\lambda)$.
2. Solve $g'(\lambda) = x - \Lambda'(\lambda) = 0$ for $\lambda^\star$ via Newton's method: $\lambda^{(k+1)} = \lambda^{(k)} - \frac{x - \Lambda'(\lambda^{(k)})}{-\Lambda''(\lambda^{(k)})}$, using $\Lambda'' > 0$ (convexity) to guarantee a well-posed Newton step.
3. $\Lambda^\star(x) = g(\lambda^\star) = \lambda^\star x - \Lambda(\lambda^\star)$.
4. If Λ is only available via simulation (e.g. $\Lambda(\lambda) = \log(\frac{1}{M}\sum_m e^{\lambda X^{(m)}})$ for samples $X^{(m)}$), Λ', Λ'' can be estimated by finite differences, or by direct simulation of the tilted moments $\mathbb{E}_{\nu_\lambda}[X], \mathrm{Var}_{\nu_\lambda}(X)$ via reweighting the same sample set with weights $e^{\lambda X^{(m)} - \Lambda(\lambda)}$.

The Newton iteration converges quadratically (typically 3-5 iterations suffice to machine precision) because Λ is smooth and strictly convex on the interior of its domain (Section 17.2) – the well-behaved convex-optimisation structure of $\Lambda^\star$ is, computationally, one of the most pleasant aspects of large deviation theory: unlike many variational problems, the relevant optimisation (a 1D convex problem, or a d-dimensional convex problem in the multivariate case) is essentially always tractable once Λ itself can be evaluated.

Key formulas: Section 17.2

- CGF: $\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}]$; Legendre transform: $\Lambda^\star(x) = \sup_\lambda(\lambda x - \Lambda(\lambda))$.
- Cramér's theorem: $\bar{X}_n$ satisfies LDP, rate n, rate function $I = \Lambda^\star$.
- Chernoff bound: $\Pr(\bar{X}_n \geq x) \leq e^{-n\Lambda^\star(x)}$ for $x > \mathbb{E}[X]$.
- Exponential tilt: $\nu_\lambda(dx) = e^{\lambda x - \Lambda(\lambda)}\nu(dx)$, $\mathbb{E}_{\nu_\lambda}[X] = \Lambda'(\lambda)$.
- Gaussian: $\Lambda^\star(x) = \frac{(x-\mu)^2}{2\sigma^2}$. Exponential($\theta$): $\Lambda^\star(x) = \theta x - 1 - \log(\theta x)$.

Notation and further reading. We write Λ for the cumulant generating function (CGF, some-

times denoted ψ or K in other texts) and $\Lambda^\star$ for its Legendre–Fenchel transform (also called the convex conjugate, Fenchel conjugate, or rate function depending on context – all the same object in this setting). The domain $\mathcal{D}_\Lambda = \{\lambda : \Lambda(\lambda) < \infty\}$ is always an interval containing 0. Readers interested in the convex-analytic foundations underlying the Legendre transform's properties (continuity, differentiability, biconjugacy $\Lambda^{\star\star} = \Lambda$ for convex lsc Λ) should consult Rockafellar's *Convex Analysis*, the standard reference for this material across optimisation, economics, and large deviation theory alike.

Cross-volume connections. The Chernoff bound and exponential tilting introduced here reappear, essentially unchanged, in Volume II's analysis of reliability-system failure-time tail bounds (where X_i might represent component lifetimes or inter-failure times) and in Volume IV's analysis of cascading failure probabilities in networks (where $\Lambda^\star$ governs the probability that a randomly-selected fraction of network nodes fail simultaneously). The convex-duality structure $(\Lambda, \Lambda^\star)$ also mirrors the duality structure central to Volume III's robust optimisation framework, where a "worst-case" (robust) objective is often expressed as a Legendre-transform-type dual of a nominal objective plus an ambiguity penalty – the same mathematical machinery (convex conjugates) underlying both large deviation rate functions and robust optimisation duality.

Example — Extended discussion: correlated sums and the multivariate CGF

For correlated Gaussian $\mathbf{X} \sim \mathcal{N}(\square, \square)$, $\Lambda(\square) = \square\cdot\square + \frac{1}{2}\square^\top \square\square$ and $\Lambda^\star(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\square)^\top \square^{-1}(\mathbf{x}-\square)$ – the multivariate Mahalanobis-distance rate function (Appendix A's multivariate extension).

Worked example: two correlated assets. Let $\square = (0, 0)$, $\sigma_1 = \sigma_2 = 0.02$, correlation $\rho_{12} = 0.6$, so $\square = \begin{pmatrix} 0.0004 & 0.00024 \\ 0.00024 & 0.0004 \end{pmatrix}$, $\det\square = 1.024 \times 10^{-7}$. For $\mathbf{x} = (-0.04, -0.04)$ (both assets down 4%), $\mathbf{x}^\top\square^{-1}\mathbf{x} = \frac{0.0016}{1.024\times10^{-7}}[0.0004-0.00048+0.0004] = \frac{0.0016\times0.00032}{1.024\times10^{-7}} \approx 5.0$, so $\Lambda^\star(\mathbf{x}) \approx 2.5$.

Comparison with independence. If $\rho_{12} = 0$, $\Lambda^\star_{\text{indep}}(\mathbf{x}) = \frac{x_1^2}{2\sigma_1^2} + \frac{x_2^2}{2\sigma_2^2} = 4.0$. Correlation *reduces* the rate function for "both fall together" from 4.0 to 2.5: a joint $4\%/4\%$ drawdown is $e^{n(4.0-2.5)} = e^{1.5n}$ times *more likely* under $\rho = 0.6$ than under independence, for n-day sums. This quantifies, via the multivariate rate function alone, the familiar risk-management intuition that positive correlation dramatically increases the probability of simultaneous large losses – the same "common factor" effect Section 20.5 explores at the aggregate-loss level.

Example — Extended discussion: heavy tails and the breakdown of exponential rate functions

Cramér's theorem requires $\Lambda(\lambda) = \mathbb{E}[e^{\lambda X}] < \infty$ in a neighbourhood of $\lambda = 0$. For heavy-tailed X with $\Pr(X > x) \sim x^{-\alpha}$, $\Lambda(\lambda) = \infty$ for all $\lambda > 0$, so $I(x) = 0$ for every $x > \mu$ – the LDP machinery degenerates because large deviations are no longer "improbable along a smooth path" but dominated by a single extreme jump.

Consequence. For such losses, $\Pr(S_n > nx)$ for $x > \mu$ is asymptotically $n\Pr(X_1 > nx)$ rather than exponential in n – the "one big jump" principle. This is precisely why catastrophe-risk tails (Section 20.5) cannot rely on Cramér-type exponential bounds and instead require the regularly-varying/subexponential machinery of Appendix F's reference distributions.

17.3 Sanov's Theorem and Relative Entropy

Overview

Sanov's theorem lifts large deviations from *numbers* (sample means) to *distributions* (empirical measures). It says that the probability an i.i.d. sample's empirical distribution $\hat{\nu}_n$ looks like some other distribution ρ decays as $e^{-nD(\rho\|\nu)}$, where $D(\cdot\|\cdot)$ is the Kullback–Leibler divergence – the rate function *is* relative entropy.

Definition: Empirical measure and relative entropy

For $X_1, \dots, X_n$ i.i.d. on a finite alphabet $\mathcal{A} = \{1, \dots, K\}$ with law $\nu = (\nu_1, \dots, \nu_K)$, the *empirical measure* is

$$\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad \hat{\nu}_n(j) = \frac{1}{n}\#\{i \le n : X_i = j\}.$$

The *Kullback–Leibler divergence* (relative entropy) between ρ, ν on $\mathcal{A}$ is

$$D(\rho\|\nu) = \sum_{j=1}^{K} \rho_j \log \frac{\rho_j}{\nu_j}, \qquad \text{with } D(\rho\|\nu) = +\infty \text{ if } \rho_j > 0 = \nu_j \text{ for some } j.$$

Properties of D. $D(\rho\|\nu) \ge 0$ with equality iff $\rho = \nu$ (Gibbs' inequality, via Jensen's inequality applied to $-\log$, a convex function); D is jointly convex in (ρ, ν); D is *not* symmetric and does not satisfy the triangle inequality – it is a divergence, not a metric, though it locally behaves like $\frac{1}{2}(\rho-\nu)^{\top}\mathbf{F}(\rho-\nu)$ for the Fisher information metric $\mathbf{F}$ (a χ^2-type quadratic form), giving it a Riemannian interpretation on the probability simplex.

Example — Theorem: Sanov's theorem (finite alphabet)

$(\hat{\nu}_n)_{n\ge1}$, viewed as random elements of the probability simplex $\Delta_K = \{\rho \in \mathbb{R}^K_{\ge 0} : \sum_j \rho_j = 1\}$, satisfies the LDP with rate n and good convex rate function $I(\rho) = D(\rho\|\nu)$:

$$\Pr(\hat{\nu}_n \approx \rho) \approx e^{-nD(\rho\|\nu)}.$$

Proof via the method of types. The number of sequences $x_1, \dots, x_n \in \mathcal{A}^n$ with empirical distribution exactly ρ (a "type") is the multinomial coefficient $\binom{n}{n\rho_1, \dots, n\rho_K}$, and each such sequence has probability $\prod_j \nu_j^{n\rho_j}$ under $\nu^{\otimes n}$. Hence

$$\Pr(\hat{\nu}_n = \rho) = \binom{n}{n\rho_1, \dots, n\rho_K} \prod_{j=1}^{K} \nu_j^{n\rho_j}.$$

By Stirling, $\frac{1}{n}\log\binom{n}{n\rho_1,\dots,n\rho_K} \to H(\rho) = -\sum_j \rho_j \log \rho_j$ (Shannon entropy of ρ), and $\frac{1}{n}\log\prod_j \nu_j^{n\rho_j} = \sum_j \rho_j \log \nu_j$. Adding,

$$\frac{1}{n}\log \Pr(\hat{\nu}_n = \rho) \to H(\rho) + \sum_j \rho_j \log \nu_j = -\sum_j \rho_j \log \frac{\rho_j}{\nu_j} = -D(\rho\|\nu).$$

The number of distinct types is only polynomial in n (at most $(n+1)^K$), so it does not affect the exponential rate when summing over ρ in a set A: $\Pr(\hat{\nu}_n \in A) \approx e^{-n \inf_{\rho\in A} D(\rho\|\nu)}$, the LDP.

Example — Worked example: a loaded die under scrutiny

A die is claimed to be fair, $\nu = (\frac{1}{6}, \dots, \frac{1}{6})$. After $n = 600$ rolls, the empirical distribution is $\hat{\nu}_{600} = (0.10, 0.10, 0.10, 0.10, 0.10, 0.50)$ – face 6 appears half the time. Compute $D(\hat{\nu}\|\nu)$:

$$D(\hat{\nu}\|\nu) = \sum_{j=1}^{5} 0.10 \log \frac{0.10}{1/6} + 0.50 \log \frac{0.50}{1/6} = 5(0.10)\log(0.6) + 0.50\log(3).$$

Numerically: $5 \times 0.10 \times (-0.5108) = -0.2554$, and $0.50 \times 1.0986 = 0.5493$, so $D(\hat{\nu}\|\nu) \approx 0.2939$ nats. Sanov's theorem says $\Pr(\hat{\nu}_{600} \approx \hat{\nu}) \approx e^{-600 \times 0.2939} \approx e^{-176.3}$ – astronomically small, quantifying just how overwhelming the evidence of loading is. Contrast with a milder deviation, $\hat{\nu}' = (0.15, 0.15, 0.15, 0.15, 0.15, 0.25)$: $D(\hat{\nu}'\|\nu) = 5(0.15)\log(0.9) + 0.25\log(1.5) \approx 5(0.15)(-0.1054) + 0.25(0.4055) \approx -0.0790 + 0.1014 = 0.0224$, giving Pr $\approx e^{-600 \times 0.0224} = e^{-13.4} \approx 1.5 \times 10^{-6}$ – still rare, but far less so, illustrating the extreme sensitivity of $D(\rho\|\nu)$ (and hence the exponential rate) to the size and direction of the deviation $\rho - \nu$.

General (non-finite) Sanov's theorem and projections. For a general Polish space $\mathcal{X}$ (not just finite alphabets), Sanov's theorem holds in the weak topology on $\mathcal{P}(\mathcal{X})$ with rate function $D(\rho\|\nu)$ for $\rho \ll \nu$ (and $+\infty$ otherwise), provided D is defined via the Radon–Nikodym derivative $D(\rho\|\nu) = \int \log \frac{d\rho}{d\nu}\, d\rho$. The Bernoulli rate function of Section 17.1, $I(x) = x\log(x/p) + (1-x)\log((1-x)/(1-p))$, is precisely $D(\text{Bern}(x)\|\text{Bern}(p))$ – Cramér's theorem for a scalar mean is the Sanov rate function *projected* onto the one-dimensional statistic $\rho \mapsto \sum_j \rho_j x_j$ (the mean), via the contraction principle of Section 17.5. This hierarchy – Sanov (distributions) $\to$ contraction $\to$ Cramér (means) – organises the whole of classical large deviation theory and explains why $D(\cdot\|\cdot)$ recurs as the canonical "cost" functional throughout robust statistics and information theory.

Example — Exercises

1. Compute $D(\rho\|\nu)$ for $\rho = (0.5, 0.5)$, $\nu = (0.9, 0.1)$. *Solution:* $D = 0.5\log(0.5/0.9) + 0.5\log(0.5/0.1) = 0.5\log(0.556) + 0.5\log 5 \approx 0.5(-0.588) + 0.5(1.609) = -0.294 + 0.805 = 0.511$ nats.

2. Compute $D(\nu\|\rho)$ for the same ρ, ν and confirm $D(\rho\|\nu) \neq D(\nu\|\rho)$. *Solution:* $D(\nu\|\rho) = 0.9\log(0.9/0.5) + 0.1\log(0.1/0.5) = 0.9\log 1.8 + 0.1\log 0.2 \approx 0.9(0.588) + 0.1(-1.609) = 0.529 - 0.161 = 0.368$. Indeed $0.368 \neq 0.511$ – D is asymmetric, as claimed; the "cost" of mistaking a fair coin for a $90/10$ coin differs from the reverse.

3. For $K = 3$, $\nu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, compute $D(\rho\|\nu)$ for $\rho = (0.5, 0.3, 0.2)$ and interpret via Sanov's theorem for $n = 300$. *Solution:* $D = 0.5\log(0.5/(1/3)) + 0.3\log(0.3/(1/3)) + 0.2\log(0.2/(1/3)) = 0.5\log 1.5 + 0.3\log 0.9 + 0.2\log 0.6 \approx 0.5(0.405) + 0.3(-0.105) + 0.2(-0.511) = 0.203 - 0.032 - 0.102 = 0.069$. So $\Pr(\hat{\nu}_{300} \approx (0.5, 0.3, 0.2)) \approx e^{-300 \times 0.069} = e^{-20.7} \approx 10^{-9}$ – a roughly 1-in-a-billion empirical deviation from a uniform die over 300 rolls, a useful benchmark for "how surprising is this deviation."

Further perspectives. Sanov's theorem (1957) was originally stated for the empirical distribution of i.i.d. samples on a finite or countable alphabet; its extension to general Polish spaces (Section 17.3's closing remarks) required substantially more sophisticated tools (weak topology on $\mathcal{P}(\mathcal{X})$, projective limit arguments) developed over the following decades. The relative entropy $D(\rho\|\nu)$ that emerges as the rate function has a dual life as the central object of information theory, where $D(\rho\|\nu)$ quantifies the expected number of extra bits needed to encode samples from ρ using a code optimised for ν – Sanov's theorem can thus be read

as: "the probability that an i.i.d. ν-sample looks like it came from ρ decays at the rate of the coding inefficiency of using a ν-optimised code for ρ-distributed data," a beautiful coding-theoretic gloss on a purely probabilistic statement. This information-theoretic connection also explains why $D(\rho\|\nu)$, rather than some other divergence (e.g. total variation, Hellinger, χ^2), is the "correct" rate function: it is the unique (up to scaling) divergence that is additive over independent observations – $D(\rho^{\otimes n}\|\nu^{\otimes n}) = nD(\rho\|\nu)$ – exactly matching the e^{-nD} scaling of the LDP. Other divergences (e.g. $\chi^2(\rho\|\nu)$) appear as the *quadratic approximation* to D near $\rho = \nu$ (via a Taylor expansion of D, giving the Fisher-information/χ^2 structure mentioned in the section), relevant for *local* (CLT-scale) asymptotics rather than large-deviation-scale asymptotics.

Example — Application note: model validation via empirical-distribution divergence

A credit rating agency's internal model assigns probabilities $\nu = (\nu_{AAA}, \dots, \nu_D)$ to 8 rating categories for a cohort of $n = 400$ bonds. After one year, the realised distribution of actual outcomes (which category each bond's rating "transitions to" or "defaults into") is $\hat{\nu}_{400}$. Sanov's theorem provides a principled test: if $D(\hat{\nu}_{400}\|\nu)$ is "large" relative to $1/400$ (i.e. $400 \times D(\hat{\nu}_{400}\|\nu)$ is large compared to, say, a χ^2-type reference), the model ν is poorly calibrated. Suppose $400 \times D(\hat{\nu}_{400}\|\nu) \approx 18.2$. Comparing to a χ^2_7 distribution (one less than the 8 categories, the standard degrees-of-freedom for a goodness-of-fit test on a 8-category multinomial) – recalling that for ρ close to ν, $2nD(\rho\|\nu) \approx n\chi^2(\rho,\nu)$ (the local quadratic approximation to relative entropy mentioned in Section 17.3's remarks, where $\chi^2(\rho,\nu) = \sum_j(\rho_j-\nu_j)^2/\nu_j$) – so $2\times 18.2 = 36.4$ would be compared to χ^2_7, whose 99% critical value is about 18.5: a value of 36.4 far exceeds this, strongly rejecting the null hypothesis that ν correctly describes the outcome distribution. This connects Sanov's theorem directly to one of the oldest and most widely used statistical tests (Pearson's χ^2 goodness-of-fit test, 1900) – the χ^2 test is, in this light, simply the large-n, local (small-deviation) approximation to a Sanov-theorem-based likelihood-ratio test, with $D(\rho\|\nu)$ the "exact" (large-deviation-valid for *any* size deviation) generalisation of $\frac{1}{2}\chi^2(\rho,\nu)$.

Algorithm: computing the Sanov rate function via convex optimisation

Input: reference distribution $\nu = (\nu_1, \dots, \nu_K)$; linear constraint $\sum_j \rho_j c_j = m$ (e.g. $c_j = x_j$ for the mean constraint of Section 17.5).

1. Form the Lagrangian $\mathcal{L}(\rho, \lambda, \mu) = \sum_j \rho_j \log(\rho_j/\nu_j) - \lambda(\sum_j \rho_j c_j - m) - \mu(\sum_j \rho_j - 1)$.
2. Setting $\partial\mathcal{L}/\partial\rho_j = 0$: $\log(\rho_j/\nu_j) + 1 - \lambda c_j - \mu = 0 \Rightarrow \rho_j = \nu_j e^{\lambda c_j + \mu - 1}$.
3. Normalisation ($\sum_j \rho_j = 1$) fixes μ in terms of λ: $\rho_j = \nu_j e^{\lambda c_j}/Z(\lambda)$, $Z(\lambda) = \sum_j \nu_j e^{\lambda c_j}$.
4. Solve $\sum_j \rho_j(\lambda) c_j = m$ (a 1D equation in λ, via the same Newton iteration as Algorithm 17.2) for $\lambda^\star$.
5. $I_y(m) = D(\rho(\lambda^\star)\|\nu) = \lambda^\star m - \log Z(\lambda^\star)$.

Step 5's formula, $I_y(m) = \lambda^\star m - \log Z(\lambda^\star)$, is recognisable as exactly $\Lambda^\star(m)$ for $\Lambda(\lambda) = \log Z(\lambda) = \log \sum_j \nu_j e^{\lambda c_j} = \log \mathbb{E}_\nu[e^{\lambda c(X)}]$ – the algorithm for Sanov's rate function under a linear constraint *is*, computationally, identical to Algorithm 17.2 for Cramér's theorem, with Λ defined via the constraint function c. This is the contraction principle (Section 17.5) made fully algorithmic: any linearly-constrained Sanov problem reduces to a Cramér-type 1D convex optimisation, regardless of the alphabet size K.

Key formulas: Section 17.3

- Empirical measure: $\hat{\nu}_n = \frac{1}{n}\sum_i \delta_{X_i}$.
- Relative entropy: $D(\rho\|\nu) = \sum_j \rho_j \log \frac{\rho_j}{\nu_j} \geq 0$, with equality iff $\rho = \nu$.
- Sanov's theorem: $\hat{\nu}_n$ satisfies LDP, rate n, rate function $I(\rho) = D(\rho\|\nu)$.
- Method of types: $\Pr(\hat{\nu}_n = \rho) = \binom{n}{n\rho_1,\ldots,n\rho_K}\prod_j \nu_j^{n\rho_j}$, giving $\frac{1}{n}\log \Pr \to H(\rho) + \sum_j \rho_j \log \nu_j = -D(\rho\|\nu)$.

Notation and further reading. The relative entropy $D(\rho\|\nu)$ is written using natural logarithms throughout this volume (units of "nats"); the information-theory literature often uses base-2 logarithms ("bits") – a constant factor $\log 2$ relates the two, with no effect on any LDP rate-function comparison (which involves only ratios or comparisons of D-values, where the constant cancels). The empirical measure $\hat{\nu}_n$ is sometimes denoted L_n ("empirical law") or P_n in other texts. Cover & Thomas's *Elements of Information Theory* provides the standard treatment of the information-theoretic side of relative entropy (coding interpretations, Fisher information, the method of types), complementing the large-deviation treatment here.

Cross-volume connections. The relative entropy $D(\rho\|\nu)$ that emerges as Sanov's rate function is the same divergence that appears as the "ambiguity measure" in Volume III's distributionally-robust optimisation (where a decision-maker optimises against the worst distribution ρ within a $D(\rho\|\nu) \leq \eta$ ball around a nominal ν) – Sanov's theorem provides the large-deviation *justification* for using D-balls as ambiguity sets: a $D(\rho\|\nu) \leq \eta$ ball is, to leading exponential order, exactly the set of empirical distributions $\hat{\nu}_n$ that are not "exponentially rare" at level $e^{-n\eta}$, giving η a precise probabilistic interpretation as a confidence-level-like parameter, directly linking this Part's pure probability theory to Volume III's decision-theoretic framework.

Example — Extended discussion: Pinsker's inequality and total variation

$D(\rho\|\nu)$ bounds total variation $\|\rho - \nu\|_{TV} = \frac{1}{2}\sum_j |\rho_j - \nu_j|$ via Pinsker's inequality: $\|\rho - \nu\|_{TV} \leq \sqrt{\frac{1}{2}D(\rho\|\nu)}$ – a one-directional bound (small D implies small TV, but not conversely).

Worked example. For $\rho = (0.5, 0.3, 0.2)$, $\nu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $D(\rho\|\nu) \approx 0.069$ nats (Section 17.3's exercises). Pinsker gives $\|\rho - \nu\|_{TV} \leq \sqrt{0.0345} \approx 0.186$. The exact $\|\rho - \nu\|_{TV} = \frac{1}{2}(|0.5 - \frac{1}{3}| + |0.3 - \frac{1}{3}| + |0.2 - \frac{1}{3}|) = \frac{1}{2}(0.1667 + 0.0333 + 0.1333) \approx 0.1667$ – indeed ≤ 0.186, with a modest $\sim 10\%$ gap.

Application. If a regulator requires "the realised default-rate distribution must be within 5% total variation of the model," Pinsker translates this into $D(\hat{\nu}_n\|\nu) \leq 2(0.05)^2 = 0.005$ as a *sufficient* condition. Sanov's theorem then gives $\Pr(D(\hat{\nu}_n\|\nu) > 0.005) \approx e^{-n\times 0.005}$ for the closest ρ with $D(\rho\|\nu) = 0.005$ – a conservative large-deviation bound for "probability the realised distribution violates the 5% TV requirement," without solving the constrained optimisation over $\{\rho : \|\rho - \nu\|_{TV} > 0.05\}$ directly.

Example — Extended discussion: the chi-squared approximation to relative entropy

For ρ close to ν (write $\rho_j = \nu_j(1 + \epsilon_j)$, $\sum_j \nu_j \epsilon_j = 0$, ϵ_j small), a second-order expansion gives $D(\rho\|\nu) \approx \frac{1}{2}\sum_j \nu_j \epsilon_j^2 = \frac{1}{2}\chi^2(\rho, \nu)$, Pearson's chi-squared statistic. Sanov's theorem then reduces, near ν, to $\Pr(D(\hat{\nu}_n\|\nu) \approx d) \approx e^{-nd} \approx e^{-n\chi^2/2}$ – recovering

classical chi-squared goodness-of-fit asymptotics as the small-deviation limit of the large-deviation rate.
Numeric check. For $\rho = (0.34, 0.33, 0.33)$, $\nu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $\chi^2 \approx 3 \times (0.01)^2/(1/3) \approx 0.0003$, so $D \approx 0.00015$ – matching a direct computation closely, confirming the quadratic approximation is excellent for "near-null" model-validation checks, unlike the large deviations from ν that drive Section 17.3's main rare-event applications.

17.4 The Gärtner–Ellis Theorem

Overview

Cramér's theorem requires i.i.d. summands. The Gärtner–Ellis theorem removes this restriction: whenever the *scaled* cumulant generating function of a (possibly dependent) sequence converges to a limit Λ, the corresponding LDP holds with rate function $\Lambda^\star$ – provided Λ is sufficiently smooth. This is the workhorse tool for Markov chains, time series, and stochastic-process functionals.

Definition: Limiting scaled CGF

Let (Z_n) be a sequence of $\mathbb{R}$-valued random variables (e.g. $Z_n = S_n$, a partial sum, with scale $a_n = n$). Define

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{a_n} \log \mathbb{E}[e^{\lambda a_n Z_n}],$$

assumed to exist (possibly $= +\infty$) for all $\lambda \in \mathbb{R}$, and to be *differentiable* on the interior of $\mathcal{D}_\Lambda = \{\lambda : \Lambda(\lambda) < \infty\}$ (the "steepness" condition at the boundary of $\mathcal{D}_\Lambda$, if finite).

Example — Theorem: Gärtner–Ellis

Under the smoothness/steepness assumptions above, (Z_n) satisfies the LDP with rate a_n and rate function

$$I(x) = \Lambda^\star(x) = \sup_{\lambda\in\mathbb{R}} (\lambda x - \Lambda(\lambda)),$$

i.e. exactly the Legendre–Fenchel transform formula of Cramér's theorem, but now Λ is the *limiting* scaled CGF of a general (dependent) sequence rather than the CGF of a single i.i.d. summand.
Why differentiability matters. The upper bound (Chernoff-type) $\frac{1}{a_n} \log \Pr(Z_n \geq x) \leq -\sup_{\lambda\geq 0}(\lambda x - \Lambda_n(\lambda))$ holds for the finite-n CGF Λ_n under essentially no assumptions (it is just Markov's inequality plus convexity, exactly as in Section 17.2). The *lower* bound requires constructing a tilted measure under which x becomes typical, and this tilting argument needs Λ to be differentiable at the optimising $\lambda^\star(x)$ so that the tilted mean $\Lambda'(\lambda^\star)$ is well-defined and equals x. When Λ has a "flat" or non-differentiable region (e.g. a phase transition in the limiting free energy), the rate function $I = \Lambda^\star$ may fail to be *strictly* convex, and the LDP can fail for sets straddling the non-differentiable point – a genuine phenomenon in statistical mechanics models exhibiting phase transitions.

Example — Worked example: additive functionals of a two-state Markov chain

Let $(Y_k)_{k\geq 0}$ be a Markov chain on $\{0,1\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \qquad \alpha,\beta \in (0,1),$$

and consider $S_n = \sum_{k=1}^{n} Y_k$ (the time spent in state 1). The scaled CGF $\Lambda(\lambda) = \lim_n \frac{1}{n}\log \mathbb{E}[e^{\lambda S_n}]$ can be computed via the Perron–Frobenius eigenvalue of the "tilted" transition matrix $\mathbf{P}_\lambda$ with entries $(\mathbf{P}_\lambda)_{ij} = P_{ij}e^{\lambda \mathbf{1}_{j=1}}$:

$$\mathbf{P}_\lambda = \begin{pmatrix} 1-\alpha & \alpha e^{\lambda} \\ \beta & (1-\beta)e^{\lambda} \end{pmatrix}, \qquad \Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda),$$

where $\rho(\cdot)$ denotes the spectral radius (Perron–Frobenius eigenvalue, which is real, positive, and simple for this irreducible non-negative matrix). Explicitly, $\rho(\mathbf{P}_\lambda)$ is the larger root of

$$\rho^2 - [(1-\alpha)+(1-\beta)e^{\lambda}]\rho + (1-\alpha-\beta)e^{\lambda} = 0.$$

For $\alpha = \beta = 0.3$ (a symmetric chain with stationary distribution $(0.5, 0.5)$ and stationary mean $\mathbb{E}_\pi[Y] = 0.5$), at $\lambda = 0$ this gives $\rho^2 - 1.4\rho + 0.4 = 0$, roots $\rho = 1, 0.4$; the Perron root is $\rho = 1$, so $\Lambda(0) = 0$ as required. The rate function $I(x) = \Lambda^\star(x)$ then governs $\Pr(S_n/n \approx x) \approx e^{-nI(x)}$ for the *time-average occupancy* of state 1 – a quantity central to reliability theory (Volume II), where Y_k might indicate "system in failed state at time k" and x a target unavailability level. The Gärtner–Ellis machinery converts a question about a dependent, correlated process into the same variational form as the i.i.d. Cramér case, but with Λ now encoding the chain's mixing structure through $\rho(\mathbf{P}_\lambda)$.

Connections. Gärtner–Ellis subsumes Cramér's theorem (i.i.d. case: $\Lambda(\lambda) = \Lambda_1(\lambda)$, the single-step CGF, independent of n) and provides the analytic backbone for Section 18.3 (LDPs for Markov chains via the Donsker–Varadhan variational formula) and for the additive functionals that appear throughout reliability and queueing applications in Volume II. The Perron–Frobenius eigenvalue formula above is the discrete-time analogue of the principal-eigenvalue formulas that govern large deviations of diffusions (Section 18.1).

Example — Exercises

1. For the symmetric chain $\alpha = \beta = 0.3$ at $\lambda = 0.5$, compute $\rho(\mathbf{P}_{0.5})$ from the quadratic $\rho^2 - [(1-\alpha)+(1-\beta)e^{0.5}]\rho + (1-\alpha-\beta)e^{0.5} = 0$. *Solution:* $(1-\alpha)+(1-\beta)e^{0.5} = 0.7 + 0.7\times 1.6487 = 0.7 + 1.1541 = 1.8541$; $(1-\alpha-\beta)e^{0.5} = 0.4 \times 1.6487 = 0.6595$. Solve $\rho^2 - 1.8541\rho + 0.6595 = 0$: $\rho = \frac{1.8541 \pm \sqrt{1.8541^2 - 4\times 0.6595}}{2} = \frac{1.8541 \pm \sqrt{3.438 - 2.638}}{2} = \frac{1.8541 \pm \sqrt{0.800}}{2} = \frac{1.8541 \pm 0.894}{2}$. The Perron root is $\rho \approx 1.374$, giving $\Lambda(0.5) = \log 1.374 \approx 0.317$.

2. Verify $\Lambda(0) = 0$ for this chain directly from the quadratic at $\lambda = 0$. *Solution:* At $\lambda = 0$, $\mathbf{P}_0 = \mathbf{P}$ (the original stochastic matrix), whose Perron eigenvalue is 1 (row sums $= 1$) – giving $\Lambda(0) = \log 1 = 0$, as required of any CGF.

3. Discuss why $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$ is convex in λ. *Solution:* $\rho(\mathbf{P}_\lambda)$ is, by the Perron–Frobenius variational formula, $\rho(\mathbf{P}_\lambda) = \sup_{\mathbf{v}>0} \min_i \frac{(\mathbf{P}_\lambda \mathbf{v})_i}{v_i}$ (or via the Collatz–Wielandt formula) – a supremum of functions that are log-linear in λ (since each entry of $\mathbf{P}_\lambda$ is $P_{ij}e^{\lambda \mathbf{1}_{j=1}}$, log-linear in λ), and a supremum of convex (here, log-convex) functions is convex – hence $\log \rho(\mathbf{P}_\lambda) = \Lambda(\lambda)$ is convex, as required for any CGF/rate-

function pair by the general theory of Section 17.2.

Further perspectives. The Gärtner–Ellis theorem (Gärtner 1977, Ellis 1984) was developed independently in the contexts of statistical mechanics (Ellis, studying Curie–Weiss-type mean-field models) and probability theory (Gärtner, studying Markov additive processes), and their convergence onto the same theorem is a striking example of the unifying power of the large-deviation formalism. The "steepness" condition on Λ – differentiability at the boundary of its domain, or $|\Lambda'(\lambda)| \to \infty$ as λ approaches the boundary of $\mathcal{D}_\Lambda$ – deserves a concrete illustration: for $\mathcal{D}_\Lambda = (-\infty, \lambda_{\max})$ with $\lambda_{\max} < \infty$ finite, steepness requires $\Lambda'(\lambda) \to \infty$ as $\lambda \to \lambda_{\max}^-$; if instead $\Lambda'(\lambda) \to L < \infty$ (a finite limit), the rate function $\Lambda^\star(x)$ becomes *linear* (not strictly convex) for $x > L$ – a "phase transition" in the rate function reflecting that the LDP upper bound remains valid but the lower bound (tilting) argument breaks down, since no finite tilt $\lambda^\star$ can produce mean exactly $x > L$. Such non-steep CGFs arise for random variables with exponential (rather than lighter) tails – e.g. $X \sim \mathrm{Exp}(\theta)$ has $\Lambda(\lambda) = -\log(1 - \lambda/\theta) \to \infty$ as $\lambda \to \theta^-$, so $\Lambda'(\lambda) \to \infty$ too (steep, no issue) – but heavier-than-exponential (subexponential) tails, central to Volume I's treatment of extreme value theory, fall *outside* the Cramér/Gärtner–Ellis framework entirely ($\Lambda(\lambda) = \infty$ for all $\lambda > 0$), requiring the entirely different "one big jump" asymptotics that resurface in Section 20.5.

Example — Application note: regime-switching volatility and VaR backtesting

A risk model assumes daily returns follow a regime-switching process: a hidden 2-state Markov chain (low-volatility regime 0, high-volatility regime 1, with the transition matrix $\mathbf{P}$ of Section 17.4's worked example, $\alpha = \beta = 0.3$) determines which of two Gaussian return distributions ($\mathcal{N}(0, \sigma_0^2)$ in regime 0, $\mathcal{N}(0, \sigma_1^2)$ in regime 1, $\sigma_1 > \sigma_0$) generates each day's return. The long-run fraction of days spent in the high-volatility regime is $\pi(1) = 0.5$ (by the chain's symmetry). A VaR backtest asks: over a rolling $n = 250$-day window, what is the probability that the *realised* fraction of high-volatility days exceeds, say, 0.65 (a 30% excess over the long-run 0.5)? This is exactly the Donsker–Varadhan / Gärtner–Ellis setup of Section 17.4 (and Section 18.3), with $f(s) = \mathbf{1}_{s=1}$ and target $x = 0.65$. Using the Perron-eigenvalue formula $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$ from the worked example, a numerical evaluation (e.g. scanning λ to find $\Lambda'(\lambda) = 0.65$, the tilt that makes 0.65 typical) might yield $I(0.65) \approx 0.05$, giving $\Pr(\text{fraction} \geq 0.65) \approx e^{-250 \times 0.05} = e^{-12.5} \approx 3.7 \times 10^{-6}$ – an extremely rare regime-composition outcome over 250 days. If a VaR backtest *observes* such an outcome (or a comparably extreme one) within a few years of monitoring, this is strong evidence against the assumed regime-switching parameters (α, β) – perhaps the true high-volatility regime is more persistent (smaller β, the probability of *exiting* the high-volatility state) than assumed, a hypothesis directly testable by re-estimating (α, β) and checking whether $I(0.65)$ under the re-estimated chain is consistent with the observed backtesting frequency – a model-calibration loop entirely powered by the Gärtner–Ellis rate function.

Algorithm: computing $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$ via power iteration

Input: transition matrix $\mathbf{P}$ (size $K \times K$), reward function f, tilt λ.

1. Form $\mathbf{P}_\lambda$ with entries $(\mathbf{P}_\lambda)_{ij} = P_{ij} e^{\lambda f(j)}$.
2. Initialise $\mathbf{v}^{(0)} = (1, \dots, 1)^\top$.
3. Iterate $\mathbf{v}^{(k+1)} = \mathbf{P}_\lambda \mathbf{v}^{(k)} / \|\mathbf{P}_\lambda \mathbf{v}^{(k)}\|$ (power iteration) until convergence; $\rho(\mathbf{P}_\lambda) \approx \|\mathbf{P}_\lambda \mathbf{v}^{(k)}\|$ at convergence (the Rayleigh-quotient-type estimate for the dominant

eigenvalue).

4. $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$.
5. To compute $I(x) = \Lambda^\star(x)$, wrap this procedure in the Newton iteration of Algorithm 17.2, with $\Lambda'(\lambda)$ estimated by finite differences (re-running the power iteration at $\lambda \pm h$ for small h) or, more efficiently, via the eigenvector: $\Lambda'(\lambda) = \frac{\mathbf{v}_\lambda^\top \frac{\partial \mathbf{P}_\lambda}{\partial \lambda} \mathbf{v}_\lambda}{\mathbf{v}_\lambda^\top \mathbf{P}_\lambda \mathbf{v}_\lambda}$ (a standard eigenvalue-perturbation formula), avoiding repeated power iterations.

Power iteration converges geometrically at a rate governed by the ratio of the second-largest to largest eigenvalue of $\mathbf{P}_\lambda$ (the matrix's "spectral gap") – for chains with strong mixing (large spectral gap, fast convergence to stationarity), $\Lambda(\lambda)$ is computed in just a handful of iterations; for slowly-mixing chains (small spectral gap, as might arise in the regime-switching application of Section 17.4's application note if regime persistence $1-\alpha, 1-\beta$ is close to 1), many more iterations are needed – the same spectral gap that governs mixing-time (and hence CLT variance, per Section 18.3's remarks) also governs the computational cost of evaluating the rate function itself, a pleasing (if perhaps unsurprising) unity between the statistical and computational aspects of the theory.

Key formulas: Section 17.4

- Limiting scaled CGF: $\Lambda(\lambda) = \lim_n \frac{1}{a_n} \log \mathbb{E}[e^{\lambda a_n Z_n}]$.
- Gärtner–Ellis: $I = \Lambda^\star$, under differentiability/steepness of Λ.
- Markov-chain additive functionals: $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$, $(\mathbf{P}_\lambda)_{ij} = P_{ij} e^{\lambda f(j)}$ (Perron–Frobenius eigenvalue).
- Non-steep $\Lambda \Rightarrow$ rate function $\Lambda^\star$ can become linear (loses strict convexity) beyond a threshold.

Notation and further reading. $\rho(\cdot)$ denotes the spectral radius (Perron–Frobenius eigenvalue for non-negative matrices), not to be confused with the use of ρ as a generic probability vector in Section 17.3 or as a correlation coefficient elsewhere in this volume – context disambiguates throughout. The "tilted matrix" $\mathbf{P}_\lambda$ construction is sometimes called the "twisted" or "exponentially tilted" transition kernel in the Markov-chain large deviations literature. For the underlying Perron–Frobenius theory, see Seneta's *Non-negative Matrices and Markov Chains*; for the original Gärtner–Ellis theorem and its many variants, see Dembo & Zeitouni, Chapter 2.3.

Cross-volume connections. The Perron–Frobenius eigenvalue construction $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$ for Markov-chain additive functionals reappears in Volume IV's analysis of network cascade dynamics, where the relevant "transition matrix" describes how failure probabilities propagate between network nodes, and the analogous tilted-matrix eigenvalue governs the exponential rate of large-scale cascade probabilities. The Gärtner–Ellis theorem's steepness condition, and its failure for non-steep CGFs (Section 17.4's remarks), directly foreshadows Volume I's treatment of heavy-tailed (regularly-varying) distributions, for which $\Lambda(\lambda) = \infty$ for all $\lambda > 0$ – the "non-steep, in fact infinite-domain" extreme case, requiring the entirely different asymptotic toolkit of regular variation theory that Volume I develops.

Example — Extended discussion: continuous-time Markov-modulated rate functions

The discrete-time chain extends to continuous time: a generator matrix $\mathbf{Q}$ ($Q_{ii} = -\sum_{j \neq i} Q_{ij}$) and a tilted generator $\mathbf{Q}_\lambda$ with $(\mathbf{Q}_\lambda)_{ii} = Q_{ii} + \lambda f(i)$ (off-diagonals un-

changed) has leading eigenvalue $\Lambda(\lambda)$ with $\Lambda(0) = 0$.

Worked example. Consider $\mathbf{Q} = \begin{pmatrix} -0.5 & 0.5 \\ 2.0 & -2.0 \end{pmatrix}$ (state 0 "normal," mean sojourn 2; state 1 "stressed," mean sojourn 0.5; stationary $\pi = (0.8, 0.2)$). With $f(0) = 0, f(1) = 1$, $\mathbf{Q}_\lambda = \begin{pmatrix} -0.5 & 0.5 \\ 2.0 & -2.0 + \lambda \end{pmatrix}$. The leading eigenvalue solves $\Lambda^2 + (2.5 - \lambda)\Lambda - 0.5\lambda = 0$. At $\lambda = 0$: $\Lambda \in \{0, -2.5\}$, $\Lambda(0) = 0$. ✓ At $\lambda = 0.5$: $\Lambda^2 + 2\Lambda - 0.25 = 0 \Rightarrow \Lambda = \frac{-2+\sqrt{5}}{2} \approx 0.118$.

Relevance. This continuous-time formulation is the natural setting for Appendix G's trading-desk model, where "days" could be replaced by exponentially-distributed regime durations – the discrete- and continuous-time formulations give qualitatively identical rate-function behaviour, with the generator-eigenvalue problem replacing the transition-matrix eigenvalue problem of Section 17.4's main treatment.

Example — Extended discussion: non-differentiability of Λ and phase transitions

The Gärtner–Ellis theorem requires $\Lambda(\lambda) = \lim_n \frac{1}{n} \log \mathbb{E}[e^{\lambda S_n}]$ to be differentiable on the relevant interval. A "kink" – a point $\lambda^\star$ where left and right derivatives differ – can cause the LDP to fail or yield only an upper bound, signalling a genuine phase transition: beyond some tilt, the dominant contribution shifts discontinuously between regimes.

Worked check. In this section's two-state model, $\Lambda(\lambda)$ (the leading eigenvalue of $\mathbf{Q}_\lambda$) remains smooth for all real λ because the chain is irreducible with a strictly positive spectral gap. Non-differentiability instead arises in models with multiple communicating classes or a continuum of metastable states, where the "effective" generator eigenvalue can switch branches. Detecting kinks in an empirically estimated $\hat{\Lambda}_n(\lambda)$ is itself a diagnostic for hidden regime structure, complementing the explicit two-state model above.

17.5 Contraction Principles and Applications

Overview

The contraction principle is the transfer mechanism of large deviation theory: if (μ_n) on $\mathcal{X}$ satisfies an LDP with rate function $I_{\mathcal{X}}$, and $f : \mathcal{X} \to \mathcal{Y}$ is continuous, then the pushforward $(\mu_n \circ f^{-1})$ on $\mathcal{Y}$ also satisfies an LDP, with rate function obtained by *minimising $I_{\mathcal{X}}$ over the fibre $f^{-1}(y)$*. It is the rigorous justification for "deriving Cramér from Sanov" and for transferring LDPs along any deterministic, continuous transformation.

Example — Theorem: Contraction principle

Let (μ_n) on $\mathcal{X}$ satisfy the LDP with good rate function $I_{\mathcal{X}}$, and let $f : \mathcal{X} \to \mathcal{Y}$ be continuous. Then $(\mu_n \circ f^{-1})$ on $\mathcal{Y}$ satisfies the LDP with good rate function

$$I_{\mathcal{Y}}(y) = \inf\{I_{\mathcal{X}}(x) : x \in \mathcal{X},\ f(x) = y\}, \qquad (\inf \emptyset := +\infty).$$

Proof idea. For the upper bound, $\{f(X) \in C\} = \{X \in f^{-1}(C)\}$ for closed C, and $f^{-1}(C)$ is closed by continuity of f, so $\limsup_n \frac{1}{a_n} \log \Pr(f(X_n) \in C) \le -\inf_{f^{-1}(C)} I_{\mathcal{X}} = -\inf_{y \in C} I_{\mathcal{Y}}(y)$ by definition of $I_{\mathcal{Y}}$. For the lower bound, given open $U \subseteq \mathcal{Y}$ and $y \in U$ with $I_{\mathcal{Y}}(y)$ close to $\inf_U I_{\mathcal{Y}}$, pick $x \in f^{-1}(y)$ with $I_{\mathcal{X}}(x)$ close to $I_{\mathcal{Y}}(y)$ (possible since the

infimum defining I_y is attained, by compactness of level sets of $I_{\mathcal{X}}$ and continuity of f); then $f^{-1}(U)$ is open and contains x, so $\liminf_n \frac{1}{a_n}\log \Pr(X_n \in f^{-1}(U)) \geq -I_{\mathcal{X}}(x) \approx -I_y(y)$. Taking y to approach the infimum gives the result.

Example — Worked example: Cramér's theorem as a contraction of Sanov's theorem

Take $\mathcal{X} = \Delta_K$ (the simplex of empirical distributions on alphabet $\{x_1, \dots, x_K\} \subset \mathbb{R}$), with Sanov's rate function $I_{\mathcal{X}}(\rho) = D(\rho\|\nu)$ (Section 17.3). Let $f : \Delta_K \to \mathbb{R}$ be the linear "sample mean" map $f(\rho) = \sum_{j=1}^{K} \rho_j x_j = \mathbb{E}_\rho[X]$, which is continuous. By the contraction principle, the empirical mean $\bar{X}_n = f(\hat{\nu}_n)$ satisfies the LDP on $\mathbb{R}$ with rate function

$$I_y(m) = \inf\Big\{ D(\rho\|\nu) : \rho \in \Delta_K,\ \textstyle\sum_j \rho_j x_j = m \Big\}.$$

This is a finite-dimensional convex optimisation: minimise relative entropy subject to a linear moment constraint. By Lagrangian duality (introducing a multiplier λ for the constraint $\sum_j \rho_j x_j = m$ and $\rho_j^\star \propto \nu_j e^{\lambda x_j}$ – an exponential tilt!), the minimum value is exactly

$$I_y(m) = \sup_{\lambda \in \mathbb{R}} \Big(\lambda m - \log \sum_{j=1}^{K} \nu_j e^{\lambda x_j} \Big) = \Lambda^\star(m),$$

recovering Cramér's theorem (Section 17.2) verbatim, with $\Lambda(\lambda) = \log \sum_j \nu_j e^{\lambda x_j} = \log \mathbb{E}_\nu[e^{\lambda X}]$ the CGF. For the Bernoulli case ($K = 2$, $x_1 = 0, x_2 = 1$, $\nu = (1-p, p)$), the constrained minimiser is $\rho^\star = (1-m, m)$ (the *only* distribution on $\{0,1\}$ with mean m), so $I_y(m) = D((1-m,m)\|(1-p,p))$, precisely the rate function of Section 17.1. The contraction principle thus reveals Cramér's theorem to be nothing but Sanov's theorem viewed through the lens $f(\rho) = \mathbb{E}_\rho[X]$, with the optimisation collapsing (for $K = 2$) to a single point because the mean constraint pins down the entire two-point distribution.

General applications and outlook. The contraction principle applies far beyond means: it transfers LDPs along any continuous statistic – variances, quantiles, ratios, solutions of equations parametrised by $\mathcal{X}$, or (in Section 18.x) entire sample paths mapped through continuous functionals such as hitting times or maxima. Whenever a quantity of interest $Y = f(X)$ is a continuous function of an object X whose large-deviation behaviour is already known, the contraction principle hands over the rate function for Y "for free," via the variational minimisation $\inf_{f^{-1}(y)} I_{\mathcal{X}}$. This is the principle that, in Part 18, allows large deviation rate functions for stochastic differential equations (path space) to induce rate functions for exit times and exit locations (a continuous functional of the path) – the Freidlin–Wentzell theory.

Part 18

Large Deviations for Stochastic Processes

Sample-path large deviations and the Freidlin–Wentzell theory for dynamical systems with noise.

Example — Exercises

1. Let $f(\rho) = \sum_j \rho_j x_j^2$ (second moment, rather than mean). State the contraction-principle formula for the rate function of $\bar{X^2}_n = \frac{1}{n}\sum_i X_i^2$. *Solution:* $I_y(m) = \inf\{D(\rho\|\nu) : \sum_j \rho_j x_j^2 = m\}$ – formally identical in structure to the mean case, but

with the linear constraint $\sum \rho_j x_j = m$ replaced by the (still linear in ρ, since x_j^2 are fixed constants) constraint $\sum \rho_j x_j^2 = m$. The dual (Lagrangian) form is $I_y(m) = \sup_\lambda(\lambda m - \log \sum_j \nu_j e^{\lambda x_j^2})$.

2. For $K = 2$, $x_1 = 0, x_2 = 1$ (Bernoulli alphabet), note $x_j^2 = x_j$, so the second-moment functional coincides with the mean functional. What does this imply about $I_y(m)$ for the second moment in this case? *Solution:* Since $x_j^2 = x_j$ for $x_j \in \{0, 1\}$, $\bar{X^2}_n = \bar{X}_n$ exactly (not just in distribution), so the rate function for the second moment is *identical* to the rate function for the mean – $I_y(m) = D((1-m, m)\|(1-p, p))$, the same Bernoulli-KL formula as Section 17.1. This degeneracy is special to $0/1$-valued alphabets and would not hold for, e.g., $x_1 = -1, x_2 = 2$.

3. Use the contraction principle informally to argue that $\max(\bar{X}_n^{(1)}, \bar{X}_n^{(2)})$ for two independent Cramér-type sequences satisfies an LDP with rate function $I_{\max}(m) = \min(I_1(m), I_2(m))$ for m exceeding both means. *Solution:* $\{\max(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}) \geq m\} = \{\bar{X}_n^{(1)} \geq m\} \cup \{\bar{X}_n^{(2)} \geq m\}$, a union of two independent rare events each with probability $\approx e^{-nI_i(m)}$; the union's probability is $\approx e^{-nI_1(m)} + e^{-nI_2(m)} \approx e^{-n \min(I_1(m), I_2(m))}$ (the larger of two exponentially small terms dominates the sum) – so $I_{\max} = \min(I_1, I_2)$, the "easier" (less costly) of the two ways to make the maximum large dominates, a simple instance of the "one big jump"-type reasoning that recurs in Section 20.5.

Further perspectives. The contraction principle, in the form stated here, requires f to be *continuous*; extensions to merely measurable or discontinuous f exist but require additional care (the inverse images $f^{-1}(C)$ of closed sets need not be closed, breaking the upper-bound argument). A practically important continuous (but non-linear) application beyond the worked example: for $\mathbf{X}_n \in \mathbb{R}^d$ satisfying an LDP with rate function $I_{\mathcal{X}}$, the contraction principle applied to $f(\mathbf{x}) = \|\mathbf{x}\|$ (any norm) gives the LDP rate function for $\|\mathbf{X}_n\|$ as $I_y(r) = \inf\{I_{\mathcal{X}}(\mathbf{x}) : \|\mathbf{x}\| = r\}$ – the infimum of $I_{\mathcal{X}}$ over a sphere, a genuinely $(d-1)$-dimensional optimisation for $d \geq 2$ that typically has no closed form but is amenable to numerical optimisation (e.g. via Lagrange multipliers for the sphere constraint), directly relevant to multivariate risk aggregation (Section 20.5) where $\|\mathbf{X}\|$ might represent a portfolio's total exposure. A second, dynamic application: the contraction principle applied to the time-T evaluation map $f(\phi) = \phi(T)$ on path space (Section 18.1) gives the LDP rate function for X_T^ϵ alone (forgetting the rest of the path) as $I_T(y) = \inf\{I_{[0,T]}(\phi) : \phi(T) = y\}$ – precisely the quasi-potential-type infimum of Section 18.2, now showing explicitly that "quasi-potential" is itself an instance of the contraction principle applied to path-space LDPs, tying together Parts 17 and 18 at the level of the underlying variational structure.

Example — Application note: contraction principle for a leveraged position

An investor holds a leveraged position $Y = L \cdot X$ for leverage ratio $L = 3$ and underlying return X satisfying Cramér's LDP with rate function $I_X(x) = \Lambda_X^\star(x)$. The map $f(x) = Lx$ is continuous (linear), so by the contraction principle, $Y = f(X)$ satisfies an LDP with rate function $I_Y(y) = \inf\{I_X(x) : Lx = y\} = I_X(y/L)$ (the infimum is trivial since f is a bijection here – a single point $x = y/L$ achieves $f(x) = y$). For $X \sim \mathcal{N}(0, \sigma^2)$, $I_X(x) = x^2/(2\sigma^2)$, so $I_Y(y) = I_X(y/L) = (y/L)^2/(2\sigma^2) = y^2/(2L^2\sigma^2)$ – exactly the rate function of $\mathcal{N}(0, L^2\sigma^2)$, as expected ($Y = LX \sim \mathcal{N}(0, L^2\sigma^2)$ exactly, for Gaussian X). The interesting case is non-Gaussian X: for X with a non-quadratic $\Lambda_X^\star$ (e.g. the asymmetric cubic-CGF example of Section 17.2's application note), $I_Y(y) = I_X(y/L)$ *rescales* the argument by $1/L$ but does not otherwise change the *shape* of the rate function – in particular, any asymmetry (different rates for gains

vs. losses) present in I_X is preserved, but evaluated at a rescaled threshold. For $L = 3$ and a loss threshold of $y = -3$ (i.e. a -3-unit loss on the leveraged position, corresponding to $x = -1$ on the underlying), $I_Y(-3) = I_X(-1)$ – the leveraged loss probability $\Pr(\bar{Y}_n \le -3) \approx e^{-nI_X(-1)}$ is, to leading exponential order, *identical* to the underlying's $\Pr(\bar{X}_n \le -1)$ probability: leverage by a constant factor L simply rescales the loss *threshold* at which a given probability is attained, by exactly L, without otherwise distorting the large-deviation rate – a clean, rigorous statement of "leverage scales losses linearly" at the level of rate functions, valid for arbitrary (non-Gaussian) underlying return distributions, as long as the contraction principle's continuity requirement (trivially satisfied for linear f) holds.

Algorithm: numerical contraction via constrained optimisation

Input: rate function $I_{\mathcal{X}}$ on $\mathcal{X}$ (e.g. given as an explicit function, or as $D(\cdot\|\nu)$ on a simplex); continuous map $f : \mathcal{X} \to \mathcal{Y}$; target $y \in \mathcal{Y}$.

1. Formulate $I_y(y) = \min_x I_{\mathcal{X}}(x)$ subject to $f(x) = y$.
2. If $I_{\mathcal{X}}$ is convex and f is affine, this is a convex program (e.g. solvable via the Lagrangian/Newton approach of Algorithm 17.3 for linear f).
3. If f is non-linear (e.g. $f(\rho) = \sum_j \rho_j x_j^2$, the second-moment functional of Section 17.5's exercises, which is linear *in* ρ even though it involves x_j^2, so still tractable by Algorithm 17.3 with $c_j = x_j^2$) – but for genuinely non-linear-in-ρ constraints (e.g. $f(\rho) = (\sum_j \rho_j x_j)^2$, the squared mean), use a general nonlinear-programming solver (e.g. sequential quadratic programming), initialised from the unconstrained minimiser of $I_{\mathcal{X}}$ (typically $\rho = \nu$, $I_{\mathcal{X}}(\nu) = 0$) and following the constraint manifold $f(x) = y$ as y varies (a "continuation" or "path-following" approach), using the solution at nearby y as a warm start.

The continuation strategy in step 3 is particularly effective because $I_y(y)$ is typically a smooth, convex function of y near $y = f(x^\star)$ (the LLN value, where $I_y = 0$) – small steps in y produce small changes in the optimal x, so warm-starting from the previous solution dramatically reduces the number of nonlinear-solver iterations needed at each new y, turning a sequence of potentially-hard nonlinear programs into a sequence of easy, well-initialised ones – the same "follow the solution as a parameter varies" idea that underlies the adaptive level-selection of Section 19.3's subset simulation and the iterative threshold schedule of Section 19.4's cross-entropy method.

Key formulas: Section 17.5

- Contraction principle: f continuous, (μ_n) LDP with good $I_{\mathcal{X}} \Rightarrow (\mu_n \circ f^{-1})$ LDP with $I_y(y) = \inf\{I_{\mathcal{X}}(x) : f(x) = y\}$.
- Cramér from Sanov: $f(\rho) = \sum_j \rho_j x_j$, $I_y(m) = \inf\{D(\rho\|\nu) : \sum_j \rho_j x_j = m\} = \Lambda^\star(m)$.
- For $0/1$-alphabet, the constrained minimiser is unique: $\rho^\star = (1-m, m)$.

Notation and further reading. The map f in the contraction principle is required only to be continuous on the relevant domain (typically a sub-level set of $I_{\mathcal{X}}$, since the infimum defining $I_{\mathcal{X}}$ is only meaningfully attained on compact sets for good rate functions) – global continuity on all of $\mathcal{X}$ is sufficient but not necessary. The contraction principle is sometimes stated as a "transfer principle" or, in the language of category theory (echoing this volume's companion MathLang series), as a statement that the LDP rate function defines a (contravariant)

functor from continuous maps to rate functions. Dembo & Zeitouni, Theorem 4.2.1, is the standard reference for the precise statement and proof.

Cross-volume connections. The contraction principle's "transfer LDPs along continuous maps" logic is the abstract version of a calculation performed concretely in Volume IV: given an LDP for individual component failure indicators, the contraction principle (applied to the map "indicator vector $\mapsto$ system functioning/failed," determined by the network's structure function) transfers the LDP to the system-level failure probability – with the infimum in the contraction formula corresponding to identifying the "minimal cut set" (the cheapest combination of component failures that causes system failure) in reliability terminology. This Section's abstract variational machinery is, in this sense, the large-deviation-theoretic restatement of classical structural reliability's minimal-cut-set analysis, now generalised to continuous (not just binary) component states.

Example — Extended discussion: rate function for a Sharpe-ratio-type statistic

Many risk metrics are ratios of sample averages, e.g. $\widehat{SR}_n = \bar{X}_n / \sqrt{\overline{X^2}_n - \bar{X}_n^2}$. The contraction principle applies via $f(m, v) = m/\sqrt{v - m^2}$ on the joint statistic $(\bar{X}_n, \overline{X^2}_n)$, but the joint LDP requires the bivariate Cramér theorem for (X_i, X_i^2) (Section 17.2's extended discussion), since $\bar{X}_n$ and $\overline{X^2}_n$ are dependent.

Qualitative behaviour. Near $s = \mu/\sigma$ (the true Sharpe ratio, where $I_{SR} = 0$), the delta-method analogue of the contraction principle gives $\mathrm{Var}(\widehat{SR}_n) \approx \frac{1}{n}\left(1 + \frac{s^2}{2}\right)$ for Gaussian X_i, so $I''_{SR}(\mu/\sigma) = 1/\left(1 + \frac{(\mu/\sigma)^2}{2}\right)$ – *larger* true Sharpe ratios correspond to *smaller* curvature, i.e. a *flatter* rate function near its minimum, i.e. *more* relative variability in *estimated* Sharpe ratios for high-Sharpe strategies. The contraction-principle framework explains this well-documented phenomenon as a direct consequence of the non-linear map $f(m, v) = m/\sqrt{v - m^2}$ having a flatter rate-function landscape away from $s = 0$.

Example — Extended discussion: contracting the rate function for VaR and Expected Shortfall

$\mathrm{VaR}_\alpha(\hat{F}_n) = \hat{F}_n^{-1}(\alpha)$ and $\mathrm{ES}_\alpha(\hat{F}_n) = \mathbb{E}[\hat{F}_n^{-1}(U) \mid U > \alpha]$ are both images of the empirical loss distribution under contraction maps. The contraction principle gives a simple quantile-shift rate function $I_{\mathrm{VaR}}(\delta) = I(F^{-1}(\alpha) + \delta)$ for VaR, but I_{ES} involves contracting over the entire tail beyond α – an infinite-dimensional contraction with no closed form in general.

Consequence. Because ES aggregates the whole tail rather than a single quantile, its rate function is generically *flatter* near the true value than VaR's – ES estimators have larger relative sampling variability for the same n, a quantitative restatement (via rate-function curvature, as in Section 17.1's moderate-deviation discussion) of the well-known elicitability/estimability trade-off between the two risk measures.

18.1 Sample-Path Large Deviations: The Freidlin–Wentzell Theory

Overview

The Freidlin–Wentzell theory extends large deviation theory from random *numbers* to random *paths*: for a diffusion with vanishingly small noise, the probability that the entire trajectory stays close to a given deterministic path ϕ decays as $e^{-I(\phi)/\epsilon}$, where $I(\phi)$ – the *action functional* – penalises deviations of $\dot\phi$ from the noiseless drift.

Definition: Small-noise diffusion and action functional

Consider the SDE on $\mathbb{R}^d$

$$dX_t^\epsilon = b(X_t^\epsilon)\,dt + \sqrt{\epsilon}\,\sigma(X_t^\epsilon)\,dW_t, \qquad X_0^\epsilon = x_0,$$

with b the drift, σ a (here, for simplicity, constant invertible) diffusion matrix, $a = \sigma\sigma^\top$, and W_t standard Brownian motion. For an absolutely continuous path $\phi : [0,T] \to \mathbb{R}^d$ with $\phi(0) = x_0$, the *Freidlin–Wentzell action functional* is

$$I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T (\dot\phi(t) - b(\phi(t)))^\top a(\phi(t))^{-1}(\dot\phi(t) - b(\phi(t)))\,dt,$$

and $I_{[0,T]}(\phi) = +\infty$ if ϕ is not absolutely continuous.

Example — Theorem: Freidlin–Wentzell LDP (Schilder's theorem as a special case)

As $\epsilon \to 0$, the law of $(X_t^\epsilon)_{t\in[0,T]}$ on the path space $C([0,T];\mathbb{R}^d)$ satisfies the LDP with rate $1/\epsilon$ and good rate function $I_{[0,T]}(\phi)$:

$$\Pr\left(X^\epsilon \approx \phi\right) \approx e^{-I_{[0,T]}(\phi)/\epsilon}.$$

Schilder's theorem is the case $b \equiv 0$, $\sigma = \mathbf{I}$: $X_t^\epsilon = \sqrt{\epsilon}\,W_t$, and

$$I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T |\dot\phi(t)|^2\,dt,$$

the Dirichlet energy of ϕ – finite only for ϕ absolutely continuous with $\dot\phi \in L^2$. This is precisely the rate function for which the most "economical" path from 0 to a target $\phi(T) = y$ in time T is the straight line $\phi(t) = yt/T$, giving $I_{[0,T]} = \frac{|y|^2}{2T}$ – consistent with the Gaussian density of $\sqrt{\epsilon}W_T \sim \mathcal{N}(0, \epsilon T\mathbf{I})$, whose log-density is $-\frac{|y|^2}{2\epsilon T}$.
Why the formula has this shape. $I_{[0,T]}(\phi)$ measures the "cost" of forcing the diffusion's drift to equal $\dot\phi$ instead of its natural drift $b(\phi)$; the discrepancy $\dot\phi - b(\phi)$ is exactly the extra "control" $u(t) = \sigma^{-1}(\dot\phi - b(\phi))$ that must be injected (via the noise channel) to realise ϕ, and $I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T |u(t)|^2 dt$ is the L^2 cost of that control – a Cameron–Martin / Girsanov change-of-measure cost. Girsanov's theorem makes this rigorous: tilting the Brownian path by drift u changes the path measure by a Radon–Nikodym factor $\exp\left(\frac{1}{\sqrt{\epsilon}}\int_0^T u\,dW_t - \frac{1}{2\epsilon}\int_0^T |u|^2 dt\right)$, whose exponential rate is exactly $-I_{[0,T]}(\phi)/\epsilon$.

Example — Worked example: small-noise gradient descent on a double-well potential

Let $b(x) = -V'(x)$ for the double-well potential $V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$, with stable equilibria at $x = \pm 1$ ($V(\pm 1) = -\frac{1}{4}$) and an unstable saddle at $x = 0$ ($V(0) = 0$). Take $\sigma = 1$, $a = 1$. The action of a path ϕ from $x_0 = -1$ to $\phi(T) = 0$ (i.e. from one well to the barrier top) is

$$I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T \left(\dot\phi + V'(\phi)\right)^2 dt.$$

A remarkable identity: expand the square,

$$\left(\dot\phi + V'(\phi)\right)^2 = \dot\phi^2 + 2\dot\phi V'(\phi) + V'(\phi)^2 = \left(\dot\phi - V'(\phi)\right)^2 + 4\dot\phi V'(\phi),$$

and $\int_0^T \dot\phi V'(\phi)\,dt = \int_0^T \frac{d}{dt}V(\phi(t))\,dt = V(\phi(T)) - V(\phi(0))$ by the chain rule. So for the *time-reversed gradient-ascent path* $\dot\phi = +V'(\phi)$ (the path that climbs the potential exactly against the natural drift), the first term vanishes and

$$I_{[0,T]}(\phi^\star) = 2\big[V(\phi(T)) - V(\phi(0))\big] = 2\big[V(0) - V(-1)\big] = 2\Big[0 - \big(-\tfrac{1}{4}\big)\Big] = \tfrac{1}{2}.$$

This is the *minimal* action over all paths from -1 to 0 (any other path incurs the extra non-negative term $(\dot\phi - V'(\phi))^2$), and it equals $2\Delta V$ where $\Delta V = V(0) - V(-1) = \frac{1}{4}$ is the barrier height. The exit probability from the well at $x = -1$ over the barrier at $x = 0$ in time T thus scales as $e^{-2\Delta V/\epsilon} = e^{-1/(2\epsilon)}$ – the celebrated *Arrhenius / Kramers law* $e^{-\Delta E/k_B T}$ from statistical physics, here with ϵ playing the role of temperature. Section 18.2 formalises $2\Delta V$ as the *quasi-potential* and Section 18.4 derives the precise (polynomial-prefactor-corrected) Eyring–Kramers exit-time formula.

Outlook. The action functional $I_{[0,T]}$ is the path-space analogue of $\Lambda^\star$ in Cramér's theorem: it is a rate function on an infinite-dimensional space (continuous paths), obtained, like Sanov's theorem, by considering an entire trajectory rather than a single summary statistic. The contraction principle (Section 17.5) then lets us push this path-space LDP forward through continuous functionals of the path – such as "first hitting time of a target set" or "value of ϕ at the first exit from a domain" – to obtain the LDPs for exit times and exit distributions studied in Sections 18.2 and 18.4.

Example — Exercises

1. For Schilder's theorem ($b \equiv 0, \sigma = 1$), verify that the straight-line path $\phi(t) = yt/T$ from 0 to y minimises $I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T \dot\phi^2 dt$ among all paths with $\phi(0) = 0, \phi(T) = y$. *Solution:* By the Cauchy–Schwarz inequality, $\left(\int_0^T \dot\phi\,dt\right)^2 \le T\int_0^T \dot\phi^2 dt$, i.e. $y^2 \le T\int_0^T \dot\phi^2 dt$, with equality iff $\dot\phi$ is constant – i.e. ϕ is the straight line. So $\int_0^T \dot\phi^2 dt \ge y^2/T$ with equality for the straight line, giving $I_{[0,T]}(\phi^\star) = \frac{1}{2}\cdot\frac{y^2}{T} = \frac{y^2}{2T}$, the minimum.
2. Confirm this matches the Gaussian density of $\sqrt{\epsilon}W_T$. *Solution:* $\sqrt{\epsilon}W_T \sim \mathcal{N}(0, \epsilon T)$, with log-density $-\frac{y^2}{2\epsilon T} + \text{const}$ – matching $-I_{[0,T]}(\phi^\star)/\epsilon = -\frac{y^2}{2T\epsilon}$ exactly, confirming Schilder's theorem reduces to the exact Gaussian tail at $\epsilon \to 0$ (here exactly, for all ϵ, since Brownian motion is exactly Gaussian).
3. For the double-well potential, compute the action of the "trivial" path that stays at the well, $\phi(t) \equiv -1$ for all $t \in [0,T]$. *Solution:* $\dot\phi \equiv 0$, $b(\phi) = -V'(-1) = 0$ (since $V'(-1) = (-1)^3 - (-1) = -1 + 1 = 0$), so $\dot\phi - b(\phi) = 0 - 0 = 0$ and $I_{[0,T]}(\phi) = 0$ –

the equilibrium path has zero action, as expected: remaining at a stable equilibrium requires no "unlikely" forcing, consistent with $I(\phi) = 0$ characterising the typical (zero-cost) behaviour.

Further perspectives. Freidlin and Wentzell's original monograph (1984, English translation) developed this theory primarily for finite-dimensional diffusions, but the action-functional formalism extends – with substantially more technical care regarding the function-space topology – to infinite-dimensional settings: stochastic partial differential equations (SPDEs), where X_t^ϵ is itself a function (e.g. a temperature or concentration field), and the action functional becomes a functional on space-time paths of fields. Such extensions underlie large-deviation analyses of, e.g., interface fluctuations in phase-separation models and rare large excursions in spatially extended ecological or epidemiological models. A second important generalisation: the constant-diffusion-matrix assumption (σ independent of x) can be relaxed to state-dependent $\sigma(x)$, in which case the action functional retains the same quadratic form $\frac{1}{2}\int(\dot\phi - b(\phi))^\top a(\phi)^{-1}(\dot\phi - b(\phi))dt$ but now with $a(\phi(t))^{-1}$ varying along the path – a "Riemannian metric" on $\mathbb{R}^d$ induced by the noise structure, under which the action functional becomes a genuine Riemannian path-length-squared functional (up to the drift term), connecting large deviation theory to sub-Riemannian and Finsler geometry in a way that has proven fruitful for understanding anisotropic noise-induced transitions (e.g. in systems where noise is much stronger in some directions than others, as is typical in biological and financial models with heterogeneous sources of randomness).

Example — Application note: action functional for an interest-rate model

The Cox–Ingersoll–Ross (CIR) interest-rate model is $dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{\epsilon r_t}\,dW_t$ (mean-reverting, with state-dependent diffusion coefficient $\sqrt{r_t}$ ensuring $r_t \geq 0$). For small ϵ (low interest-rate volatility relative to mean-reversion strength), the Freidlin–Wentzell action functional for a path ϕ is $I_{[0,T]}(\phi) = \frac{1}{2\sigma^2}\int_0^T \frac{(\dot\phi(t) - \kappa(\theta - \phi(t)))^2}{\phi(t)}\,dt$ (using $a(\phi) = \sigma^2\phi$, the CIR diffusion coefficient, in the general action functional formula of Definition (Small-noise diffusion and action functional)). Consider the question: what is the action cost for the rate to deviate from its long-run mean θ to a level $\phi(T) = \theta + \Delta$ over a horizon T, then return? For a simple "spike" path $\phi(t) = \theta + \Delta\sin(\pi t/T)$ (rising from θ to $\theta + \Delta$ at $t = T/2$ and returning to θ at $t = T$), $\dot\phi(t) = \frac{\pi\Delta}{T}\cos(\pi t/T)$, and the action integral, while not generally closed-form, can be evaluated numerically; for $\kappa = 0.5, \theta = 0.03, \sigma = 0.1, \Delta = 0.02, T = 1$ (a 2% spike in a 1-year window for a rate model with 3% long-run mean and moderate mean-reversion), numerical integration gives $I_{[0,1]}(\phi) \approx 0.18$. The interpretation: $\Pr(\text{rate path resembles this spike}) \approx e^{-0.18/\epsilon}$ – for $\epsilon = 0.01$ (a specific calibration of the model's noise intensity), this is $e^{-18} \approx 1.5 \times 10^{-8}$, an extremely rare path. This kind of action-functional evaluation underlies "scenario probability" assessments in interest-rate risk management: rather than asking "what is the probability the rate exceeds 5% at some point" (a level-crossing question, addressed by quasi-potentials, Section 18.2), one can ask "what is the probability the rate follows *this specific stress-test trajectory*" – a question the action functional answers directly, and which underlies the design of historically-motivated stress scenarios whose plausibility can be quantified via $I_{[0,T]}(\phi)$.

Algorithm: minimum action path via gradient descent on path space

Input: drift b, diffusion matrix $a = \sigma\sigma^\top$, endpoints $\phi(0) = x_0, \phi(T) = x_1$, discretisation into K time steps $\phi_0, \dots, \phi_K$ (with $\phi_0 = x_0, \phi_K = x_1$ fixed).

1. Discretise the action: $I_{[0,T]}(\phi) \approx \frac{\Delta t}{2} \sum_{k=0}^{K-1} (\dot\phi_k - b(\phi_k))^\top a(\phi_k)^{-1} (\dot\phi_k - b(\phi_k))$, with $\dot\phi_k \approx (\phi_{k+1} - \phi_k)/\Delta t$.
2. Initialise $\phi_k^{(0)}$ as the linear interpolation between x_0 and x_1.
3. Compute $\partial I / \partial \phi_k$ for each interior k (via finite differences, or analytically if b, a are smooth), and update $\phi_k \leftarrow \phi_k - \eta\, \partial I / \partial \phi_k$ (gradient descent), for a step size η.
4. Iterate until convergence (small relative change in $I_{[0,T]}(\phi)$).

This "discretise then optimise" approach – representing the path by its values at K time points and minimising the discretised action over these $K-1$ free interior points via standard nonlinear optimisation – is the computational workhorse for evaluating quasi-potentials (Section 18.2) and action functionals in non-gradient or high-dimensional systems where no closed-form formula (such as $V = 2\Delta U$ for gradient systems) is available. The free endpoint time T in the definition of the quasi-potential ($V(x) = \inf_{T,\phi} I_{[0,T]}(\phi)$) is typically handled by an outer loop over T (or by reparametrising the path on $[0,1]$ and treating T itself as an additional optimisation variable), since for many systems $I_{[0,T]}(\phi_T^\star)$ is itself a smooth, often unimodal, function of T with a well-defined minimiser $T^\star$.

Key formulas: Section 18.1

- SDE: $dX_t^\epsilon = b(X_t^\epsilon)dt + \sqrt{\epsilon}\,\sigma(X_t^\epsilon)dW_t$, $a = \sigma\sigma^\top$.
- Action functional: $I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T (\dot\phi - b(\phi))^\top a(\phi)^{-1}(\dot\phi - b(\phi))\, dt$.
- Freidlin–Wentzell LDP: $\Pr(X^\epsilon \approx \phi) \approx e^{-I_{[0,T]}(\phi)/\epsilon}$.
- Schilder's theorem ($b = 0, \sigma = \mathbf{I}$): $I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T |\dot\phi|^2 dt$; minimal path is the straight line, $I = \frac{|y|^2}{2T}$.
- Gradient identity: $\int_0^T \dot\phi V'(\phi) dt = V(\phi(T)) - V(\phi(0))$, used to find minimal-action paths.

Notation and further reading. The action functional $I_{[0,T]}$ is sometimes written $S_{[0,T]}$ (for "action," echoing the classical-mechanics Lagrangian action) in the physics literature, and I_T or simply I (with the time interval implicit) in some probability texts. The path space $C([0,T];\mathbb{R}^d)$ is endowed with the uniform (sup-norm) topology for the LDP statement. Freidlin & Wentzell's *Random Perturbations of Dynamical Systems* remains the foundational reference; Dembo & Zeitouni, Chapter 5, provides a more streamlined modern treatment of Schilder's theorem and its extensions.

Cross-volume connections. The action functional's "cost of forcing a non-typical drift" interpretation directly parallels Volume III's robust control formulations, where a decision-maker's worst-case adversary is modelled as choosing a drift perturbation $u(t)$ at L^2 cost $\frac{1}{2}\int |u|^2 dt$ – exactly the Freidlin–Wentzell action for additive noise (Schilder's theorem). This is not a coincidence: H_∞ robust control theory and large deviation theory for small-noise diffusions are, at a formal level, dual descriptions of the same min-max/variational structure, a connection made rigorous in the stochastic-control literature (and alluded to again in Section 19.2's remarks on importance sampling as stochastic control).

Example — Extended discussion: anisotropic noise and the geometry of $I_{[0,T]}$

When $a(x) = \sigma(x)\sigma(x)^\top$ is anisotropic, $I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T (\dot\phi - b(\phi))^\top a(\phi)^{-1}(\dot\phi - b(\phi))dt$ defines a non-Euclidean "cost of deviation": directions with small a are expensive to traverse, directions with large a are cheap.

Worked example. Let $b = (0,0)$, $a = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$ (noise in y is $4\times$ the variance of noise in x). Then $I_{[0,T]}(\phi) = \frac{1}{2T}(x_1^2 + \frac{1}{4}y_1^2)$ for a straight-line path to (x_1, y_1). Target A $= (1,0)$: $I = \frac{1}{2T}$. Target B $= (0,1)$: $I = \frac{1}{8T}$ – four times smaller, despite both being at Euclidean distance 1. The level sets of $I_{[0,T]}$ are ellipses $x^2 + y^2/4 =$const, elongated in the high-noise direction.

Practical relevance. If x is a low-volatility structural factor and y a high-volatility one, "rare joint scenarios" reached primarily through y are systematically *less rare* than scenarios of the same Euclidean magnitude reached through x. A stress test calibrated to fixed Euclidean "size" across factors thus implicitly over-weights low-volatility-factor stresses relative to their true rarity – an anisotropy correction the action functional makes precise.

Example — Extended discussion: sample-path LDPs for jump-diffusions

When $b(x)$, $a(x)$ are supplemented by jumps with intensity $\lambda(x)$ and jump law $\mu(x,\cdot)$, the Freidlin–Wentzell action acquires a second term: alongside the diffusive cost, each path discontinuity $\Delta\phi_i$ contributes a Cramér-type "jump cost" $I_{\text{jump}}(\Delta\phi_i) = \sup_\theta[\theta\Delta\phi_i - \lambda(\theta\mathbb{E}_\mu[e^{\theta\Delta\phi_i}] - \lambda)]$.

Consequence. When jump costs are cheaper than diffusive costs for a given displacement (heavy-tailed jump sizes, as in Section 17.2's "one big jump" discussion), the optimal escape path is no longer a smooth minimiser but a path with one or a few large jumps – the action-functional analogue of "one big jump", relevant whenever risk factors such as credit spreads or commodity prices carry jump risk on top of diffusive noise.

18.2 Quasi-Potentials and Metastability

Overview

The *quasi-potential* $V(x)$ generalises the potential-difference computation of Section 18.1 to systems whose drift b is not a pure gradient: it is the minimal action required for the noiseless-in-the-limit system to travel from a stable equilibrium to any point x, against its natural drift. It governs the exponential rate of escape from metastable states – the mathematical core of "metastability."

Definition: Quasi-potential

Let $x^\star$ be an asymptotically stable equilibrium of $\dot x = b(x)$ (i.e. a local minimum of the dynamics' basin of attraction). The *quasi-potential* relative to $x^\star$ is

$$V(x) = \inf\Big\{ I_{[0,T]}(\phi) : T > 0,\ \phi(0) = x^\star,\ \phi(T) = x \Big\},$$

the minimal Freidlin–Wentzell action (Section 18.1) over all finite-time paths from $x^\star$ to x and all durations T.

Gradient case. If $b = -\nabla U$ for a potential U (so $a = \mathbf{I}$ for simplicity), the identity used in Section 18.1's worked example generalises: for the time-reversed path $\dot\phi = +\nabla U(\phi)$ from $x^\star$ to x,

$$I_{[0,T]}(\phi) = 2\big[U(x) - U(x^\star)\big],$$

and this is the minimum (any deviation adds a non-negative term), so $V(x) = 2[U(x) - U(x^\star)]$ – the quasi-potential is simply twice the potential difference. In the *non-gradient* case (general b, e.g. systems with rotational/circulating components), V need not equal $2U$ for any single function U, and computing V requires genuinely solving the variational problem – typically via the *Hamilton–Jacobi equation* satisfied by V:

$$\tfrac{1}{2}|\nabla V(x)|^2_{a^{-1}} + b(x)\cdot\nabla V(x) = 0, \qquad V(x^\star) = 0,$$

(here $|\mathbf{y}|^2_{a^{-1}} = \mathbf{y}^\top a^{-1}\mathbf{y}$), obtained from the Euler–Lagrange equations of the action functional via the Hamiltonian formulation of the calculus of variations.

Example — Theorem: Freidlin–Wentzell exit-time asymptotics (Kramers' law)

Let D be a domain containing $x^\star$ with ∂D in the basin boundary, and let $\tau^\epsilon = \inf\{t : X^\epsilon_t \notin D\}$ be the first exit time from D. Then, as $\epsilon \to 0$,

$$\lim_{\epsilon\to 0} \epsilon \log \mathbb{E}[\tau^\epsilon] = \min_{x\in\partial D} V(x) =: V^\star,$$

i.e. $\mathbb{E}[\tau^\epsilon] \approx e^{V^\star/\epsilon}$: the mean exit time grows *exponentially* in $1/\epsilon$, with exponent equal to the minimal quasi-potential on the boundary – the height of the lowest "pass" out of the basin. Moreover, the exit location concentrates (as $\epsilon \to 0$) near the minimiser $x^\dagger = \arg\min_{\partial D} V(x)$: the system almost always escapes through the lowest pass, even though the deterministic dynamics never would.

Example — Worked example: asymmetric double-well and preferential escape route

Consider $b(x,y) = -\nabla U(x,y)$ on $\mathbb{R}^2$ with $U(x,y) = \frac{1}{4}(x^2-1)^2 + \frac{1}{2}\gamma y^2$ for $\gamma > 0$ – a "valley" along $x \in \{-1,1\}$ (two wells at $(\pm 1, 0)$) with quadratic confinement in y. The saddle connecting the wells is at $(0,0)$ with $U(0,0) = \frac{1}{4}$. By the gradient formula, $V(x,y) = 2[U(x,y) - U(-1,0)] = 2U(x,y)$ since $U(-1,0) = 0$. The minimal quasi-potential on the separatrix $\{x = 0\}$ is $V(0,y) = 2[\frac{1}{4}(0-1)^2 + \frac{1}{2}\gamma y^2] = \frac{1}{2} + \gamma y^2$, minimised at $y = 0$ giving $V^\star = \frac{1}{2}$. So $\mathbb{E}[\tau^\epsilon] \approx e^{(1/2)/\epsilon}$ – escape occurs, with overwhelming probability as $\epsilon \to 0$, through the point $(0,0)$ rather than through any $(0,y)$ with $y \neq 0$, even though the latter are also on the separatrix: the quadratic penalty γy^2 exponentially suppresses excursions away from $y = 0$. This "selection of the optimal escape route" – here trivial by symmetry, but in general the outcome of a genuine variational computation – is the geometric content of the Freidlin–Wentzell theory and underlies, e.g., the prediction of preferred failure paths in multi-component reliability networks (Volume II, Part on cascading failures).

Relation to the Eyring–Kramers prefactor. The exponential rate $V^\star = \min_{\partial D} V$ captures the leading-order behaviour, $\log \mathbb{E}[\tau^\epsilon] \approx V^\star/\epsilon$. The full *Eyring–Kramers formula* (Section 18.4) refines this to $\mathbb{E}[\tau^\epsilon] \approx C\, e^{V^\star/\epsilon}$ with an explicit constant C depending on the Hessians of U at the starting well and at the saddle – a "polynomial correction" analogous to the Stirling prefactor that large deviation rate functions alone do not capture, but which matters quantitatively for finite (not infinitesimal) noise levels.

Example — Exercises

1. For $U(x,y) = \frac{1}{4}(x^2-1)^2 + \frac{1}{2}\gamma y^2$ with $\gamma = 2$, compute $V(0, 0.5)$ using the gradient formula. *Solution:* $V(x,y) = 2U(x,y)$ (since $U(-1,0) = 0$). $U(0,0.5) = \frac{1}{4}(0-1)^2 + \frac{1}{2}(2)(0.25) = \frac{1}{4} + 0.25 = 0.5$, so $V(0,0.5) = 1.0$.
2. Compare $V(0,0.5) = 1.0$ with $V(0,0) = 2U(0,0) = 2 \times 0.25 = 0.5$, and interpret in terms of preferential escape routes. *Solution:* $V(0,0.5) = 1.0 > V(0,0) = 0.5$ – escaping via $(0,0.5)$ costs twice as much "action" as escaping via $(0,0)$, confirming that $(0,0)$ (the saddle, $y=0$) is the preferred escape point, consistent with the worked example's conclusion that escape concentrates at $y = 0$.
3. For the 1D double well $U(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$, the quasi-potential from $x^\star = -1$ satisfies $V(x) = 2[U(x) - U(-1)] = 2U(x) + \frac{1}{2}$. Compute $V(1)$ (the quasi-potential to reach the *other* well) and interpret. *Solution:* $U(1) = \frac{1}{4} - \frac{1}{2} = -\frac{1}{4}$, so $V(1) = 2(-\frac{1}{4}) + \frac{1}{2} = -\frac{1}{2} + \frac{1}{2} = 0$. $V(1) = 0$ reflects that, having paid the cost $V(0) = \frac{1}{2}$ to cross the saddle, the remainder of the journey to the other well is "downhill" (zero additional cost) – the total minimal action from -1 to 1 equals the action to reach the *highest* point along the way (the saddle), not the sum of costs to each subsequent point, consistent with V being defined via an infimum over entire paths.

Further perspectives. The Hamilton–Jacobi equation for the quasi-potential V is, formally, the zero-viscosity (or "WKB") limit of the backward Kolmogorov equation for the diffusion's transition density: writing $p_\epsilon(x,t) \sim e^{-W(x,t)/\epsilon}$ (a WKB ansatz) and substituting into the Fokker–Planck equation, the leading-order (ϵ^0) term yields exactly the Hamilton–Jacobi equation for W, with V recovered as the steady-state ($t \to \infty$) limit of W. This WKB connection is the rigorous bridge between the probabilistic (Freidlin–Wentzell) and PDE/spectral (eigenvalue-based) approaches to metastability, and explains why quasi-potential calculations and "ground-state" eigenvalue/eigenfunction calculations for the corresponding Schrödinger-type operator (obtained from the generator via a similarity transformation) give the same exponential rates – the two are dual descriptions of the same underlying variational structure. For non-gradient systems, solving the Hamilton–Jacobi equation $\frac{1}{2}|\nabla V|^2_{a^{-1}} + b \cdot \nabla V = 0$ is, in general, as hard as the original problem; however, an important class of "detailed-balance-violating but still solvable" systems exists (e.g. certain models with $b = -\nabla U + \boldsymbol{\ell}$ where $\boldsymbol{\ell}$ is a divergence-free "circulation" field orthogonal to ∇U in a suitable sense), for which $V = 2U$ still holds despite $b \neq -\nabla U$ – an important caveat to the blanket statement that non-gradient systems require genuinely new computation: *some* structure beyond pure gradient flow is compatible with the simple $V = 2U$ formula, and identifying such structure is often the key practical step in applying this theory to real (rarely purely-gradient) systems.

Example — Application note: quasi-potential for a leveraged fund's net asset value

A leveraged fund's net asset value (NAV) X_t (relative to its initial value, $X_0 = 1$) follows approximately $dX_t = \mu X_t\, dt + \sqrt{\epsilon}\sigma X_t\, dW_t$ (geometric, i.e. proportional drift and noise) for small positive μ (the fund's expected excess return) and noise scale ϵ. In log-coordinates $Y_t = \log X_t$, this becomes $dY_t = (\mu - \frac{1}{2}\epsilon\sigma^2)dt + \sqrt{\epsilon}\sigma\, dW_t$ – additive noise in Y, so the action functional (Schilder-type, Section 18.1) for Y is $I_{[0,T]}(\psi) = \frac{1}{2\sigma^2}\int_0^T (\dot\psi - (\mu - \frac{1}{2}\epsilon\sigma^2))^2 dt \approx \frac{1}{2\sigma^2}\int_0^T (\dot\psi - \mu)^2 dt$ for small ϵ (dropping the $O(\epsilon)$ correction). The "quasi-potential" question – what is the cost for the fund's log-NAV to fall to a "liquidation trigger" level $Y = -c$ (i.e. NAV falls to e^{-c} of initial value) starting from $Y_0 = 0$ – has, for this *driftless-in-the-tilted-sense* additive-noise problem,

the Schilder-type answer $V(-c) = \inf_T \inf_{\psi(0)=0,\psi(T)=-c} \frac{1}{2\sigma^2} \int_0^T (\dot{\psi} - \mu)^2 dt$. Unlike the pure Schilder case (Section 18.1, $\mu = 0$), here the drift $\mu > 0$ pushes *away* from the liquidation trigger (assuming $c > 0$, the trigger represents a *loss*), so reaching $-c$ requires "fighting" the drift – and the optimal strategy is to do so as efficiently as possible. The minimal-action path turns out to be the constant-velocity path $\dot{\psi} = -\mu$ for an appropriately chosen T (or, taking $T \to \infty$ with $\dot{\psi} = -\mu - v$ for small additional downward velocity v, optimised), giving $V(-c) \to \frac{2\mu c}{\sigma^2}$ in the relevant limit (a standard result for additive-noise processes with constant adverse drift: the quasi-potential is linear in the target distance c, unlike the quadratic dependence of the pure-diffusion Schilder case) – so $\Pr(\text{liquidation}) \approx e^{-2\mu c/(\epsilon\sigma^2)}$, decreasing *exponentially* in both the drift-to-volatility ratio μ/σ^2 (the fund's "Sharpe-ratio-like" quality) and the buffer c (distance to the liquidation trigger, in log terms) – a clean, quantitative link between fund quality, capital buffer, and survival probability.

Algorithm: numerically solving the quasi-potential Hamilton–Jacobi equation

Input: drift b, diffusion matrix a, equilibrium $x^\star$.

1. Discretise $\mathcal{X}$ on a grid; the HJ equation $\frac{1}{2}|\nabla V|^2_{a^{-1}} + b \cdot \nabla V = 0$, $V(x^\star) = 0$, is a first-order, generally non-smooth (viscosity-solution) PDE.
2. Use the *ordered upwind method* or *fast marching method* (originally developed for the Eikonal equation, which the HJ equation for V resembles closely in the isotropic case $a = \mathbf{I}$): propagate V outward from $x^\star$ in order of increasing V-value, using upwind finite-difference approximations of ∇V that respect the "causality" of the HJ equation (a grid point's V value depends only on neighbours with *smaller* V, analogous to Dijkstra's shortest-path algorithm).
3. For anisotropic a (the general case), the upwind directions must be determined by the local geometry of a^{-1}, requiring a generalisation of the standard fast-marching stencil (the "ordered upwind method" of Sethian & Vladimirsky, 2003, designed precisely for this anisotropic case).

The connection to Dijkstra's algorithm / shortest paths is not superficial: the HJ equation for V is, in the isotropic case, exactly the Eikonal equation $|\nabla V| =$ (local "slowness," determined by b and a), whose solution $V(x)$ is the *geodesic distance* from $x^\star$ to x in a Riemannian metric determined by a^{-1} – the quasi-potential is literally a distance function in an appropriate (generally non-Euclidean, and for non-gradient systems, non-symmetric "Finsler") geometry, and fast, robust algorithms for computing distance functions on grids (developed originally for image processing and robotics path-planning) transfer directly to quasi-potential computation, a nice example of methods crossing from one applied field to another via a shared underlying mathematical structure.

Key formulas: Section 18.2

- Quasi-potential: $V(x) = \inf\{I_{[0,T]}(\phi) : T > 0, \phi(0) = x^\star, \phi(T) = x\}$.
- Gradient case ($b = -\nabla U$): $V(x) = 2[U(x) - U(x^\star)]$.
- Hamilton–Jacobi equation: $\frac{1}{2}|\nabla V|^2_{a^{-1}} + b \cdot \nabla V = 0$, $V(x^\star) = 0$.
- Freidlin–Wentzell exit law: $\lim_{\epsilon\to 0} \epsilon \log \mathbb{E}[\tau^\epsilon] = \min_{\partial D} V = V^\star$; exit location concentrates at $\arg\min_{\partial D} V$.

Notation and further reading. The quasi-potential V is sometimes called the "rate function for the invariant measure" (since, under additional conditions, $\pi_\epsilon(x) \sim e^{-V(x)/\epsilon}$ for the

diffusion's stationary density π_ϵ, connecting V directly to equilibrium statistical mechanics free energies) or the "Freidlin–Wentzell potential." The notation $x^\dagger$ for the relevant saddle point is not universal; some texts use $z^\star$ or x_s. For the Hamilton–Jacobi / viscosity-solution perspective on V, see Fleming & Soner's *Controlled Markov Processes and Viscosity Solutions*, which develops the connection to stochastic control mentioned in Section 19.2's remarks.

Cross-volume connections. The quasi-potential V and its role in governing transition rates between metastable states is the continuous-state analogue of Volume II's discrete-state reliability analysis, where a system transitions between "operational" and "failed" states (or among multiple degradation states) according to transition rates that, in the small-noise/rare-transition limit, are governed by exactly the Eyring–Kramers-type formulas of Section 18.4. Volume VI's treatment of "antifragility" – systems that benefit from volatility – can be partly understood through the quasi-potential lens: a system is "antifragile" in a region of state space where increasing ϵ (volatility) *increases* $V^\star$ relative to alternatives (e.g. by widening a beneficial basin's effective quasi-potential landscape), the precise opposite of the noise-induced-tipping vulnerability discussed in Section 18.5.

Example — Extended discussion: a non-gradient system where $V = 2U$ still holds

Section 18.2 noted non-gradient systems generally require solving the HJ equation directly. Here, $b = -\nabla U + \gamma \mathbf{r}$ for $U = \frac{1}{2}(x^2 + y^2)$ and $\mathbf{r} = (-y, x)$ (a rotational field). Since $\mathbf{r} \perp \nabla U$ everywhere ($\nabla U \cdot \mathbf{r} = x(-y) + yx = 0$), $V = 2U$ *still holds* despite $b \neq -\nabla U$.

HJ verification. $\nabla V = (2x, 2y)$, $|\nabla V|^2 = 4(x^2 + y^2)$. $b \cdot \nabla V = (-x, -y) \cdot (2x, 2y) + \gamma(-y, x) \cdot (2x, 2y) = -2(x^2 + y^2) + 0$. So $\frac{1}{2}|\nabla V|^2 + b \cdot \nabla V = 2(x^2 + y^2) - 2(x^2 + y^2) = 0$ – satisfied for *any* γ.

What changes instead. The linearised dynamics near the well have eigenvalues $-1 \pm i\gamma$ – the system spirals into the well for $\gamma \neq 0$, affecting autocorrelation structure (oscillatory rather than monotonic decay, relevant to Section 18.5's early-warning indicators) and Eyring–Kramers prefactors, even though the leading-order quasi-potential $V = 2U$ – and hence the escape-rate exponent – is unaffected. This dissociation between "quasi-potential" and "local dynamics near the well" is an important subtlety whenever non-gradient drift is present.

Example — Extended discussion: entropy production and irreversibility

For non-gradient drift $b = -\nabla U + \mathbf{r}$ with $\mathbf{r} \cdot \nabla U \neq 0$ in general, the steady-state probability current $\mathbf{J} = b\pi - \frac{1}{2}a\nabla\pi$ is non-zero and the system sits in a genuine non-equilibrium steady state. The entropy production rate $\sigma = \int \frac{\mathbf{J}^\top a^{-1} \mathbf{J}}{\pi}\, dx \geq 0$ quantifies the irreversibility, vanishing exactly under detailed balance (the gradient case).

Relevance to quasi-potentials. In a NESS, V generally differs from $2U$ for any U, and Eyring–Kramers prefactors acquire $O(\sigma)$ corrections. For weakly non-gradient perturbations $b = -\nabla U + \gamma \mathbf{r}_0$, $V = 2U + O(\gamma^2)$ – consistent with the exact γ-independence found in this section's rotational-field example – but higher-order corrections appear for generic, non-orthogonal perturbations.

18.3 Large Deviations for Markov Chains

Overview

For Markov chains and processes, large deviation rate functions for additive functionals (Section 17.4, Gärtner–Ellis) admit a beautiful alternative representation due to Donsker and Varadhan: a variational formula expressing the rate function as a relative-entropy-type functional over the space of probability measures, with the Markov generator playing the role that the CGF played in Cramér's theorem.

Definition: Donsker–Varadhan rate function

Let $(Y_k)_{k\geq 0}$ be an ergodic Markov chain on a state space $\mathcal{S}$ with unique stationary distribution π, and let $f : \mathcal{S} \to \mathbb{R}$ be a bounded "reward" function. For the empirical mean $\bar{f}_n = \frac{1}{n}\sum_{k=1}^{n} f(Y_k)$, the Donsker–Varadhan rate function is

$$I(x) = \inf\Big\{ H(\mu \,|\, \mu P) : \mu \in \mathcal{P}(\mathcal{S}),\ \int f\, d\mu = x \Big\},$$

where P is the transition kernel, μP denotes μ pushed forward by one step of P, and $H(\mu \,|\, \mu P)$ is the specific relative entropy (per-step KL divergence) of the chain's actual one-step transitions under μ relative to those consistent with stationarity under μ.

Equivalence with Gärtner–Ellis. By the Donsker–Varadhan variational principle for the principal eigenvalue,

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\pi\big[e^{\lambda \sum_{k=1}^{n} f(Y_k)}\big] = \log \rho(\mathbf{P}_\lambda), \qquad (\mathbf{P}_\lambda)_{ij} = P_{ij} e^{\lambda f(j)}$$

(exactly the Perron–Frobenius construction of Section 17.4's worked example, here for a general state space), and $I(x) = \sup_\lambda (\lambda x - \Lambda(\lambda)) = \Lambda^\star(x)$, the Legendre transform – so the two formulas, one a "primal" infimum over measures and one a "dual" supremum over tilting parameters λ, give the *same* rate function $I = \Lambda^\star$, related by the same convex duality as in Cramér's theorem.

Example — Theorem: LDP for additive functionals of Markov chains

Under standard ergodicity/mixing conditions (e.g. Doeblin's condition, or geometric ergodicity for general state spaces), $(\bar{f}_n)_{n\geq 1}$ satisfies the LDP with rate n and good convex rate function $I = \Lambda^\star$ as above, where $I(\mathbb{E}_\pi[f]) = 0$ (zero cost at the stationary mean) and I is strictly convex wherever Λ is differentiable. The proof combines the Gärtner–Ellis theorem (Section 17.4), applied to $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$, with the Perron–Frobenius theorem (smoothness of $\rho(\mathbf{P}_\lambda)$ in λ, for finite or compact state spaces, follows from simplicity of the leading eigenvalue and analytic perturbation theory).

Example — Worked example: queue occupancy large deviations

Consider a discrete-time Markov chain on $\mathcal{S} = \{0, 1, \dots, N\}$ modelling a queue length (an M/M/1/N-type chain): at each step, the queue increases by 1 with probability α (arrival), decreases by 1 with probability β (departure, if queue > 0), and stays the same otherwise, with $\alpha < \beta$ (stable queue, $\rho = \alpha/\beta < 1$). Take $f(s) = \mathbf{1}_{s=N}$ (indicator of the queue being *full*). The stationary probability $\pi(N) \sim (1-\rho)\rho^N$ for large N (geometric tail, a classical M/M/1 result) already exhibits an exponential-in-N structure; the Donsker–Varadhan rate function for $\bar{f}_n = \frac{1}{n}\#\{k \leq n : Y_k = N\}$ (the

long-run *fraction of time the queue is full*) governs, for a *fixed* large N, the probability that this fraction deviates from its stationary value $\pi(N)$. For instance, with $\rho = 0.5$ and $N = 10$, $\pi(10) \approx 0.5 \times 0.5^{10} \approx 4.9 \times 10^{-4}$; the rate function $I(x)$ for $x \gg \pi(10)$ (e.g. the queue is full 5% of the time, a $\sim 100\times$ excess over stationary) would be computed from $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$ with $(\mathbf{P}_\lambda)_{ij} = P_{ij} e^{\lambda \mathbf{1}_{j=N}}$ – numerically, $\Lambda(\lambda)$ is the log of the Perron eigenvalue of $\mathbf{P}$ with the N-th column scaled by e^λ, a $(N+1) \times (N+1)$ matrix whose leading eigenvalue can be tracked as λ increases from 0. This connects directly to Volume II's reliability-theory treatment of queueing systems: "probability the system is overloaded $x\%$ of the time" is precisely a Donsker–Varadhan large deviation question.

Connection to Sanov's theorem. For i.i.d. sequences (the degenerate Markov chain with $P_{ij} = \nu_j$ independent of i), $\mu P = \nu$ for any μ, so $H(\mu \,|\, \mu P) = H(\mu \,|\, \nu) = D(\mu \| \nu)$ reduces to ordinary relative entropy, and the Donsker–Varadhan formula collapses to exactly the constrained-minimisation form of Sanov's theorem via the contraction principle (Section 17.5) – the Markov-chain LDP is the natural non-i.i.d. generalisation of the Sanov→contraction→Cramér hierarchy, with $H(\mu \,|\, \mu P)$ playing the role of $D(\rho \| \nu)$ for processes with memory.

Example — Exercises

1. For the queue chain with $\rho = 0.5, N = 10$, the stationary probability is $\pi(10) \approx (1-\rho)\rho^{10} = 0.5 \times 0.5^{10} \approx 4.88 \times 10^{-4}$. If $N = 20$ instead, compute $\pi(20)$ and compare. *Solution:* $\pi(20) \approx 0.5 \times 0.5^{20} \approx 0.5 \times 9.54 \times 10^{-7} \approx 4.77 \times 10^{-7}$ – about $1000\times$ smaller than $\pi(10)$, reflecting the geometric (ρ^N) decay: each additional unit of queue capacity multiplies the full-queue probability by $\rho = 0.5$.
2. For an i.i.d. sequence (degenerate Markov chain, $P_{ij} = \nu_j$), verify $H(\mu \,|\, \mu P) = D(\mu \| \nu)$ directly from the definition. *Solution:* $\mu P = \nu$ for *any* μ (since each row of P equals ν, the pushforward of any distribution under P is ν). The specific relative entropy $H(\mu \,|\, \mu P)$, defined via the one-step transition KL divergence, reduces (for the degenerate i.i.d. kernel) to $D(\mu \| \nu)$ directly – the "transition" KL divergence between μ's implied one-step law and ν is just $D(\mu \| \nu)$ itself, since the chain has no memory.
3. Argue informally why $I(\mathbb{E}_\pi[f]) = 0$ for the Donsker–Varadhan rate function. *Solution:* Take $\mu = \pi$ (the stationary distribution) in the infimum defining I. Then $\mu P = \pi P = \pi$ (stationarity), so $H(\pi \,|\, \pi P) = H(\pi \,|\, \pi) = 0$ (relative entropy of a distribution to itself is zero), and $\int f \, d\pi = \mathbb{E}_\pi[f]$ by definition. So $\mu = \pi$ achieves $H(\mu \,|\, \mu P) = 0$ subject to $\int f \, d\mu = \mathbb{E}_\pi[f]$ – the infimum is ≤ 0, and since $H \geq 0$ always, $I(\mathbb{E}_\pi[f]) = 0$.

Further perspectives. The Donsker–Varadhan theory (1975–1983, in a series of papers) was originally developed in the context of Markov *processes* (continuous time), where the rate function $H(\mu \,|\, \mu P)$ has a continuous-time analogue $H(\mu) = -\int \sqrt{\frac{d\mu}{d\pi}} \, \mathcal{L} \sqrt{\frac{d\mu}{d\pi}} \, d\pi$ for generator $\mathcal{L}$ (a Dirichlet-form / quadratic-form expression, for reversible processes) – a remarkably explicit formula that, for reversible diffusions, connects directly to the spectral gap of $\mathcal{L}$ and hence to mixing times, linking large deviation rates for time-averages to the same spectral quantities that govern convergence to stationarity. This spectral connection means that, for reversible Markov chains/processes, the Donsker–Varadhan rate function $I(x)$ near $x = \mathbb{E}_\pi[f]$ has curvature $I''(\mathbb{E}_\pi[f])$ directly related to the asymptotic variance $\sigma_f^2 = \lim_n n \operatorname{Var}(\bar{f}_n)$ (the "CLT variance" for the additive functional) via $I''(\mathbb{E}_\pi[f]) = 1/\sigma_f^2$ – the large-deviation rate function's local (quadratic) behaviour near its zero recovers exactly the central limit theorem variance, a beautiful instance of the general principle that LDPs "contain" CLTs as their local, second-order approximation (Bahadur–Rao-type refinements

make this correspondence precise to all orders, providing not just the leading exponential rate but polynomial-in-n correction factors analogous to the Eyring–Kramers prefactors of Section 18.4).

Example — Application note: Donsker–Varadhan for operational risk loss-event frequency

An operational risk model represents the occurrence of loss events at a financial institution as a Markov-modulated process: a hidden 2-state "control environment" chain (state 0 = "well-controlled," state 1 = "control breakdown," with the same $\mathbf{P}$ as Section 17.4's chain) modulates the daily probability of a loss event, $q_0 = 0.001$ in state 0 and $q_1 = 0.02$ in state 1 (a $20\times$ elevation during control breakdowns). The long-run (stationary) average daily loss-event probability is $\bar{q} = \pi(0)q_0 + \pi(1)q_1 = 0.5(0.001) + 0.5(0.02) = 0.0105$. Over a $n = 250$-day year, the Donsker–Varadhan rate function (applied to the additive functional $f =$ "loss event occurred," combining the regime-modulation of Section 17.4's worked example with the Bernoulli-type outcome of Section 17.1) governs $\Pr(\text{annual loss-event count}/250 \geq x)$ for $x > \bar{q}$. For $x = 0.02$ (i.e. the annual rate matches the *permanent-breakdown* rate q_1, despite the control environment being in breakdown only 50% of the time on average), the rate function $I(0.02)$ – combining both the cost of the control environment spending more time than typical in state 1 *and* the cost of even state-0 days producing more loss events than their q_0 rate would suggest – would be computed via the full Donsker–Varadhan variational formula over the joint (hidden-state, observed-event) process. While the exact numerical value requires solving the relevant Perron-eigenvalue problem for the joint chain, the *qualitative* lesson is structural: $I(0.02)$ will generally be *smaller* than the rate function for the same target under a (wrongly) assumed i.i.d. Bernoulli$(\bar{q})$ model – because the hidden Markov structure provides an "easier" (lower-cost) route to a high annual count, namely "spend more time than usual in state 1," rather than requiring every single day (even state-0 days) to be individually unlucky. This is the operational-risk analogue of the credit-portfolio correlation discussion in Section 17.1's application note: hidden common factors (here, the control-environment regime) make extreme aggregate outcomes systematically more likely than naive independence-based rate functions predict, and the Donsker–Varadhan framework is precisely the tool that incorporates this correctly.

Algorithm: empirical Donsker–Varadhan rate function estimation

Input: long simulated trajectory $(Y_0, \dots, Y_N)$ of the Markov chain, reward f.

1. For a grid of λ values, compute $\hat{\Lambda}(\lambda) = \frac{1}{n}\log\left(\frac{1}{N-n+1}\sum_{k=0}^{N-n} e^{\lambda\sum_{i=k+1}^{k+n} f(Y_i)}\right)$ for a chosen "window length" n (large enough that the sum-CGF has approached its limiting form, but small enough that $N-n+1$ (the number of overlapping windows) remains large for a stable empirical average).
2. Repeat for several n (e.g. $n = 50, 100, 200$) and check that $\hat{\Lambda}(\lambda)$ has stabilised (is approximately independent of n) – if not, increase n further (the chain's mixing time provides a guide for how large n must be).
3. Apply the Newton iteration of Algorithm 17.2 to the stabilised $\hat{\Lambda}(\lambda)$ to obtain $\hat{I}(x) = \hat{\Lambda}^\star(x)$.

The "overlapping windows" approach in step 1 makes efficient use of a single long trajectory (as opposed to simulating many independent short trajectories), but introduces dependence between the window-sums (consecutive windows share most of their data) – this depen-

dence does not bias $\hat{\Lambda}(\lambda)$ (it remains a consistent estimator as $N \to \infty$ for fixed n), but it does affect the *variance* of $\hat{\Lambda}(\lambda)$, which should be estimated via block-bootstrap or similar dependence-aware resampling methods (rather than naive i.i.d. formulas) if confidence intervals for $\hat{I}(x)$ are required – a practical instance of the general principle that time-series data require dependence-aware uncertainty quantification, distinct from (but building on) the i.i.d. formulas that suffice for Algorithms 17.1–17.3.

Key formulas: Section 18.3

- Donsker–Varadhan rate function: $I(x) = \inf\{H(\mu \,|\, \mu P) : \mu \in \mathcal{P}(\mathcal{S}), \int f \, d\mu = x\}$.
- Equivalence with Gärtner–Ellis: $I = \Lambda^\star$ for $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$.
- $I(\mathbb{E}_\pi[f]) = 0$ (achieved at $\mu = \pi$, since $H(\pi \,|\, \pi P) = H(\pi \,|\, \pi) = 0$).
- Local curvature: $I''(\mathbb{E}_\pi[f]) = 1/\sigma_f^2$ where σ_f^2 is the CLT asymptotic variance.

Notation and further reading. The "specific relative entropy" $H(\mu \,|\, \mu P)$ is also written $h(\mu)$ or $\mathcal{H}(\mu)$ in various texts, and for continuous-time processes becomes a Dirichlet form $\mathcal{E}(\sqrt{d\mu/d\pi}, \sqrt{d\mu/d\pi})$ for the generator's associated quadratic form. The original Donsker–Varadhan papers (1975a,b, 1976, 1983, in Communications on Pure and Applied Mathematics) develop the theory in considerable generality; Deuschel & Stroock's *Large Deviations* provides a more accessible modern treatment of this material alongside the Freidlin–Wentzell theory of Part 18.

Cross-volume connections. The Donsker–Varadhan rate function for Markov-chain additive functionals is the direct mathematical ancestor of Volume II's analysis of long-run system availability (the fraction of time a repairable system is operational, modelled as an additive functional of its operational/failed-state Markov chain) – "probability the system is unavailable more than $x\%$ of the time over a long horizon" is precisely a Donsker–Varadhan large deviation question, with $I(x)$ quantifying how surprising a given availability shortfall is. The spectral-gap connection to CLT variance (Section 18.3's remarks) also connects to Volume II's reliability-growth models, where the rate of convergence to a steady-state failure rate is governed by the same spectral quantities.

Example — Extended discussion: the Feynman–Kac connection for diffusions

For continuous-time diffusions, the Donsker–Varadhan analogue is the principal eigenvalue of a tilted generator $\mathcal{L} + \lambda f$, exactly the object solved (implicitly) by the Feynman–Kac formula $u(x,t) = \mathbb{E}_x[e^{\int_0^t f(X_s)ds}] \sim e^{\Lambda(\lambda)t}\psi(x)$ as $t \to \infty$.
Worked sketch. For OU process $dX_t = -X_t dt + dW_t$ (stationary $\mathcal{N}(0, 1/2)$) and $f(x) = \mathbf{1}_{x>0}$ ($\mathbb{E}_\pi[f] = 1/2$ by symmetry), the eigenvalue problem $-x\psi' + \frac{1}{2}\psi'' + \lambda \mathbf{1}_{x>0}\psi = \Lambda(\lambda)\psi$ is a Schrödinger-type problem after the standard similarity transformation. For small λ, perturbation theory gives $\Lambda(\lambda) \approx \lambda \mathbb{E}_\pi[f] + \frac{1}{2}\lambda^2\sigma_f^2$ where $\sigma_f^2 = 2\int_0^\infty \mathrm{Cov}(f(X_0), f(X_t))dt$ is the long-run variance of the time-average – recovering the CLT correspondence $I''(\mathbb{E}_\pi[f]) = 1/\sigma_f^2$ of Section 18.3's remarks, here computable explicitly from the OU process's known correlation structure $\mathrm{Corr}(X_0, X_t) = e^{-t}$, giving a complete (if computationally involved) continuous-time analogue of Section 18.3's discrete-time results.

Example — Extended discussion: level-2 and level-3 large deviations

The rate function $I(x)$ for the time-average $\bar{f}_n$ (this section's "level-1" object) is itself a contraction of a richer "level-2" rate function $H(\cdot\|\pi)$ for the empirical measure $L_n =$

$\frac{1}{n}\sum \delta_{X_i}$, given by $H(\rho\|\pi) = \sum_x \rho(x)\,\mathrm{KL}(P_\rho(\cdot|x)\,\|\,P(\cdot|x))$ for a tilted kernel P_ρ with stationary law ρ – a Markov-chain generalisation of Sanov's theorem.
Payoff. Level-3 governs the empirical *process* of consecutive-state pairs, with level-2 and level-1 both contractions of it. Any functional of long-run chain behaviour – not only simple time-averages – thus inherits a rate function via contraction from level-2/3, without a bespoke Gärtner–Ellis-type calculation each time, extending the contraction philosophy of Section 17.5 to the full Markovian setting.

18.4 Exit Times from Domains of Attraction

Overview

Section 18.2 established that $\mathbb{E}[\tau^\epsilon] \approx e^{V^\star/\epsilon}$ for the exit time τ^ϵ from a domain D, with exponential rate $V^\star = \min_{\partial D} V$. The *Eyring–Kramers formula* sharpens this to a precise asymptotic $\mathbb{E}[\tau^\epsilon] \sim C\, e^{V^\star/\epsilon}$, with C an explicit constant built from the Hessians of the potential at the starting well and at the relevant saddle point – the "prefactor" that large-deviation rate functions alone cannot see, but which is essential for quantitative predictions at realistic (non-infinitesimal) noise levels.

Definition: Hessian curvatures at well and saddle

Let $U : \mathbb{R}^d \to \mathbb{R}$ be the potential ($b = -\nabla U$, gradient case), with a local minimum at $x^\star$ (the well) and the lowest saddle point on ∂D at $x^\dagger$. Write $\mathbf{H}_\star = \nabla^2 U(x^\star)$ (positive definite, all eigenvalues > 0) and $\mathbf{H}_\dagger = \nabla^2 U(x^\dagger)$, which by definition of "saddle" has exactly one negative eigenvalue $-\mu_\dagger < 0$ (the unstable direction of steepest descent away from the saddle) and $d-1$ positive eigenvalues.

Example — Theorem: Eyring–Kramers formula

For an overdamped diffusion $dX_t = -\nabla U(X_t)dt + \sqrt{2\epsilon}\,dW_t$ (note the $\sqrt{2\epsilon}$ normalisation, standard in this context) with a unique lowest saddle $x^\dagger$ on ∂D,

$$\mathbb{E}_{x^\star}[\tau^\epsilon] \sim \frac{2\pi}{\mu_\dagger}\sqrt{\frac{|\det \mathbf{H}_\dagger|}{\det \mathbf{H}_\star}}\; e^{[U(x^\dagger)-U(x^\star)]/\epsilon}, \qquad \epsilon \to 0,$$

where $|\det \mathbf{H}_\dagger|$ denotes the absolute value of the determinant (the single negative eigenvalue contributes a sign that is absorbed). Comparing exponents: $U(x^\dagger) - U(x^\star) = \frac{1}{2}V(x^\dagger) = \frac{1}{2}V^\star$ from the gradient quasi-potential formula of Section 18.2 (note the factor-of-two difference arises from the $\sqrt{2\epsilon}$ vs. $\sqrt{\epsilon}$ noise normalisation convention) – the leading-order Freidlin–Wentzell rate is recovered, now dressed with the explicit prefactor.
One-dimensional case (*most transparent*). For $d = 1$, $\mathbf{H}_\star = U''(x^\star) > 0$ and $\mathbf{H}_\dagger = U''(x^\dagger) < 0$ (write $U''(x^\dagger) = -\mu_\dagger$), and the formula reduces to

$$\mathbb{E}_{x^\star}[\tau^\epsilon] \sim \frac{2\pi}{\sqrt{U''(x^\star)\,|U''(x^\dagger)|}}\; e^{[U(x^\dagger)-U(x^\star)]/\epsilon}.$$

Example — Worked example: quartic double well, full Eyring–Kramers calculation

Return to $U(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$ (Section 18.1), well at $x^\star = -1$ ($U(-1) = -\frac{1}{4}$), saddle at $x^\dagger = 0$ ($U(0) = 0$). Compute:

$$U''(x) = 3x^2 - 1 \;\Rightarrow\; U''(-1) = 2, \quad U''(0) = -1,$$

so $\mu_\dagger = 1$. The barrier height is $\Delta U = U(0) - U(-1) = \frac{1}{4}$. Plugging in,

$$\mathbb{E}_{-1}[\tau^\epsilon] \sim \frac{2\pi}{\sqrt{2 \times 1}}\, e^{(1/4)/\epsilon} = \pi\sqrt{2}\, e^{1/(4\epsilon)}.$$

For $\epsilon = 0.05$: $e^{1/(4\times 0.05)} = e^5 \approx 148.4$, so $\mathbb{E}[\tau^\epsilon] \approx \pi\sqrt{2} \times 148.4 \approx 659.4$ (in units of the natural diffusive timescale). Compare with $\epsilon = 0.02$: $e^{1/0.08} = e^{12.5} \approx 2.68 \times 10^5$, giving $\mathbb{E}[\tau^\epsilon] \approx 4.44 \times 2.68 \times 10^5 \approx 1.19 \times 10^6$ – a $\sim 1800\times$ increase in mean escape time from a $2.5\times$ reduction in noise intensity, vividly illustrating the exponential sensitivity that makes rare-event simulation (Part 19) necessary: direct Monte Carlo at $\epsilon = 0.02$ would require on the order of 10^6 samples just to observe a handful of transitions.

Caveats and extensions. The Eyring–Kramers formula as stated assumes a single dominant saddle; if several saddles have comparable heights $U(x_i^\dagger) - U(x^\star)$, the exit rate is a sum of contributions $\sum_i C_i e^{-\Delta U_i/\epsilon}$, and for finite ϵ the "competition" between near-degenerate saddles can produce non-trivial exit-location statistics not captured by the leading order alone. In non-gradient systems (general b, no potential U), an analogous prefactor formula exists but requires solving the Hamilton–Jacobi equation for V and analysing the linearisation of the associated Hamiltonian flow at the saddle – considerably more involved, and an active area of applied large-deviation research relevant to non-equilibrium statistical mechanics and climate tipping-point models (Section 18.5).

Example — Exercises

1. For $U(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 + 0.1x$ (a slightly tilted double well), assuming $U''(x^\star) \approx 2.05$ and $|U''(x^\dagger)| \approx 0.95$ near the original values, recompute the Eyring–Kramers prefactor $\frac{2\pi}{\sqrt{U''(x^\star)|U''(x^\dagger)|}}$. *Solution:* $\sqrt{2.05 \times 0.95} = \sqrt{1.9475} \approx 1.395$, so prefactor $\approx 2\pi/1.395 \approx 4.505$ – close to the untilted value $\pi\sqrt{2} \approx 4.443$, confirming the prefactor is relatively insensitive to small tilts, unlike the exponential term $e^{\Delta U/\epsilon}$, which is highly sensitive to ΔU.

2. For $\epsilon = 0.1$ (a noisier regime), compute $\mathbb{E}[\tau^\epsilon]$ for the untilted double well ($\Delta U = 1/4$). *Solution:* $e^{(1/4)/0.1} = e^{2.5} \approx 12.18$, so $\mathbb{E}[\tau^{0.1}] \approx \pi\sqrt{2} \times 12.18 \approx 54.1$ – a much shorter escape time than at $\epsilon = 0.05$ (≈ 659) or $\epsilon = 0.02$ ($\approx 1.19 \times 10^6$).

3. Compute the ratio $\mathbb{E}[\tau^{0.05}]/\mathbb{E}[\tau^{0.1}]$ and compare with the ratio of exponential factors alone. *Solution:* $\mathbb{E}[\tau^{0.05}]/\mathbb{E}[\tau^{0.1}] \approx 659.4/54.1 \approx 12.2$. The exponential factors alone give $e^{(1/4)/0.05}/e^{(1/4)/0.1} = e^5/e^{2.5} = e^{2.5} \approx 12.18$ – matching almost exactly, since the prefactor $\pi\sqrt{2}$ is the same constant in both, confirming that for this model essentially *all* of the ϵ-dependence is captured by the exponential term.

Further perspectives. The Eyring–Kramers formula's history traces to Arrhenius's 1889 empirical observation that chemical reaction rates scale as $e^{-E_a/k_B T}$ (the "Arrhenius law"), Kramers' 1940 derivation from the Fokker–Planck equation (the "Kramers problem," originally posed for nuclear fission rates), and Eyring's contemporaneous transition-state-theory derivation from statistical mechanics – three independent routes converging on the same prefactor formula, a convergence later placed on rigorous footing within the Freidlin–Wentzell

framework only in the 1990s–2000s (Bovier, Eckhoff, Gayrard, Klein, and others), using potential-theoretic (capacity-based) methods rather than the formal WKB/saddle-point arguments of the original derivations. The rigorous proofs revealed subtleties invisible to the formal calculation: e.g. the formula as stated requires the saddle $x^\dagger$ to be a *non-degenerate* critical point (Hessian with exactly one negative eigenvalue, none zero); at degenerate saddles (e.g. where the potential has a quartic rather than quadratic cross-section in the unstable direction), the prefactor's ϵ-dependence changes (different power of ϵ multiplies the exponential), a phenomenon relevant whenever a system is tuned exactly to a bifurcation point (Section 18.5) – precisely where the standard Eyring–Kramers formula's non-degeneracy assumption is most likely to fail, requiring the more delicate "degenerate Kramers" asymptotics for accurate prefactors near tipping points.

Example — Application note: Eyring–Kramers for inventory management

A just-in-time inventory system maintains stock level X_t governed (in a fluid-limit approximation) by $dX_t = -\nabla U(X_t)dt + \sqrt{2\epsilon}\,dW_t$ for a potential U with a stable equilibrium at the target stock level $x^\star$ and a "stockout" boundary at $x^\dagger < x^\star$ (zero inventory) modelled as a saddle in an extended (e.g. reflected or absorbing-at-zero, approximated locally by a saddle for the purposes of this estimate) potential landscape. The Eyring–Kramers formula gives $\mathbb{E}[\tau^\epsilon] \sim \frac{2\pi}{\sqrt{U''(x^\star)|U''(x^\dagger)|}} e^{[U(x^\dagger)-U(x^\star)]/\epsilon}$ for the mean time to stockout. Suppose a supply-chain redesign doubles the "restoring force" curvature $U''(x^\star)$ (e.g. via faster replenishment, tightening the effective potential well around the target level) while leaving $U(x^\dagger) - U(x^\star)$ (the "buffer," related to the target stock level itself) unchanged. The prefactor changes by a factor $1/\sqrt{2} \approx 0.707$ – a modest $\sim 30\%$ *reduction* in mean time to stockout from the prefactor alone, which might seem counterproductive (faster replenishment making stockouts *more* frequent?). The resolution: in practice, faster replenishment is typically accompanied by an opportunity to *reduce the target stock level* $x^\star$ itself (since the system can respond faster to depletion, less buffer is needed) – but reducing $x^\star$ (holding $x^\dagger = 0$ fixed) reduces the barrier $U(x^\dagger) - U(x^\star)$, and since this appears in the *exponent*, even a modest reduction in buffer can far outweigh the prefactor effect, *decreasing* $\mathbb{E}[\tau^\epsilon]$ overall (more frequent stockouts) unless the buffer reduction is calibrated carefully. This is a cautionary illustration of the Eyring–Kramers formula's two-part structure (exponential term dominates, but the prefactor is not always negligible for "moderate" parameter changes) and of the danger of optimising one part of a system (replenishment speed) without re-examining how it interacts with another (buffer sizing) through the *exponential* sensitivity of escape times to barrier height.

Algorithm: numerical Eyring–Kramers via Hessian computation

Input: potential $U(x)$ (possibly only available via numerical evaluation), well $x^\star$, saddle $x^\dagger$ (found via, e.g., gradient-based optimisation for $x^\star$ and a saddle-finding method such as the "dimer method" or "nudged elastic band" for $x^\dagger$, both standard in computational chemistry for analogous "transition state" finding problems).

1. Compute Hessians $\mathbf{H}_\star = \nabla^2 U(x^\star)$, $\mathbf{H}_\dagger = \nabla^2 U(x^\dagger)$ via finite differences (or automatic differentiation, if U is implemented in a differentiable framework).
2. Compute eigenvalues of $\mathbf{H}_\star$ (all positive, by construction of $x^\star$ as a minimum) and $\mathbf{H}_\dagger$ (exactly one negative, $-\mu_\dagger$, by construction of $x^\dagger$ as a simple saddle).
3. Evaluate the prefactor $\frac{2\pi}{\mu_\dagger}\sqrt{\frac{|\det \mathbf{H}_\dagger|}{\det \mathbf{H}_\star}}$ and the exponent $U(x^\dagger) - U(x^\star)$.

4. Combine: $\mathbb{E}[\tau^\epsilon] \approx \text{prefactor} \times e^{[U(x^\dagger)-U(x^\star)]/\epsilon}$.

The "transition-state finding" sub-problem (locating $x^\dagger$) is, in general, harder than locating $x^\star$ (a straightforward minimisation): saddle points are not local extrema of any simple merit function, and dedicated algorithms (dimer method, nudged elastic band, eigenvector-following methods) are needed, exploiting the characteristic "one negative eigenvalue" structure of $\mathbf{H}_\dagger$ to design search directions that climb in the unstable direction while descending in all stable directions simultaneously. This computational chemistry heritage – the Eyring–Kramers formula originated in chemical reaction rate theory, and the numerical methods for finding $x^\dagger$ were developed primarily for computing reaction rates of molecular conformational changes – is a striking instance of methods migrating, largely unchanged in their mathematical core, from molecular simulation to the financial, ecological, and engineering applications of Sections 18.4–18.5.

Key formulas: Section 18.4

- Eyring–Kramers: $\mathbb{E}_{x^\star}[\tau^\epsilon] \sim \frac{2\pi}{\mu_\dagger}\sqrt{\frac{|\det \mathbf{H}_\dagger|}{\det \mathbf{H}_\star}}\, e^{[U(x^\dagger)-U(x^\star)]/\epsilon}$ (overdamped, $\sqrt{2\epsilon}$ noise).
- 1D case: $\mathbb{E}[\tau^\epsilon] \sim \frac{2\pi}{\sqrt{U''(x^\star)|U''(x^\dagger)|}} e^{\Delta U/\epsilon}$.
- Quartic double-well example: $\Delta U = 1/4$, $U''(\pm 1) = 2$, $U''(0) = -1$, $\mathbb{E}[\tau^\epsilon] \sim \pi\sqrt{2}\, e^{1/(4\epsilon)}$.

Notation and further reading. The Eyring–Kramers formula is sometimes written with D (diffusion coefficient) in place of ϵ, with the precise correspondence $\epsilon \leftrightarrow D$ depending on the SDE's normalisation convention – readers consulting the physics literature should check the SDE's noise normalisation ($\sqrt{2\epsilon}$, $\sqrt{\epsilon}$, or $\sqrt{2D}$ are all common) before comparing formulas directly. Bovier & den Hollander's *Metastability: A Potential-Theoretic Approach* provides the rigorous modern treatment via capacities, including the degenerate-saddle extensions mentioned in Section 18.4's remarks.

Cross-volume connections. The Eyring–Kramers prefactor's sensitivity to Hessian curvatures at wells and saddles parallels Volume IV's analysis of network resilience, where "how robust is this network configuration to perturbation" is often quantified via analogous curvature/sensitivity measures of a network's "potential landscape" (e.g. algebraic connectivity, related to the Laplacian spectrum, playing a role analogous to $U''(x^\star)$). The chemistry/biology origins of the Eyring–Kramers formula (noted in Section 18.4's remarks) also connect directly to the mathematical-biology themes of this volume's sister series (Mathematics of Evolution and Life): conformational transition rates in protein folding, and transition rates between phenotypic states in gene-regulatory bistable systems (briefly mentioned in Section 18.5), are governed by precisely this formula.

Example — Extended discussion: a fully worked 2D Eyring–Kramers calculation

Let $U(x, y) = \frac{1}{4}x^4 - \frac{1}{2}x^2 + \frac{1}{2}cy^2$ ($c > 0$): a double well in x with transverse confinement in y. Wells at $(\pm 1, 0)$, saddle at $(0, 0)$.

Hessians. At the well $(-1, 0)$: $\mathbf{H}_\star = \begin{pmatrix} 2 & 0 \\ 0 & c \end{pmatrix}$, $\det \mathbf{H}_\star = 2c$. At the saddle $(0, 0)$: $\mathbf{H}_\dagger = \begin{pmatrix} -1 & 0 \\ 0 & c \end{pmatrix}$, $\mu_\dagger = 1$, $|\det \mathbf{H}_\dagger| = c$.

Prefactor. $\frac{2\pi}{\mu_\dagger}\sqrt{\frac{|\det \mathbf{H}_\dagger|}{\det \mathbf{H}_\star}} = 2\pi\sqrt{\frac{c}{2c}} = \pi\sqrt{2}$ – *independent of* c. This reflects exact separabil-

ity $U(x,y) = U_x(x) + U_y(y)$: the transverse confinement cancels between $\det \mathbf{H}_\star$ and $|\det \mathbf{H}_\dagger|$, reducing exactly to Section 18.4's 1D result.
Non-separable case. Adding a coupling $\delta x^2 y^2$ gives $U_{yy} = c + 2\delta$ at the well but $U_{yy} = c$ at the saddle (since $x = 0$ there), giving prefactor $\pi\sqrt{2} \times \sqrt{c/(c+2\delta)}$ – now genuinely dependent on both c and δ. Even a small non-separable coupling reintroduces transverse-direction dependence, a reminder that Algorithm 18.4's full numerical Hessian computation, not separability shortcuts, should generally be trusted for realistic potentials.

Example — Extended discussion: the Arrhenius law and noise-strength scaling

Writing the noise strength as ϵ, the Eyring–Kramers formula $\mathbb{E}[\tau] \sim Ce^{2\Delta U/\epsilon}$ is the finance analogue of the Arrhenius law, with ΔU the barrier height. Doubling ϵ does *not* halve $\log \mathbb{E}[\tau]$: $\log \mathbb{E}[\tau]$ scales as $1/\epsilon$, not linearly in ϵ.
Numeric illustration. For $\Delta U = 1$: $\epsilon = 0.1$ gives exponent 20, $\mathbb{E}[\tau] \sim e^{20} \approx 4.9 \times 10^8$; $\epsilon = 0.2$ gives exponent 10, $\mathbb{E}[\tau] \sim e^{10} \approx 2.2 \times 10^4$ – doubling the noise reduces $\mathbb{E}[\tau]$ by a factor of over 20,000. This extreme sensitivity means modest volatility misestimation can produce orders-of-magnitude errors in estimated times-to-default or times-to-breach, a caution accompanying every use of this section's main formula.

18.5 Applications to Noise-Induced Transitions

Overview

The Freidlin–Wentzell / Eyring–Kramers machinery of Sections 18.1–18.4 is not merely an abstract exercise: it is the quantitative engine behind *noise-induced transitions* – sudden, large, qualitative changes of state driven by the accumulation of small random perturbations – across physics, climate science, ecology, neuroscience, and engineered systems. This section surveys applications and connects them to the reliability and risk themes of Volumes I–II.

Climate tipping points. Paleoclimate records show abrupt transitions between climate states (e.g. Dansgaard–Oeschger events, glacial-interglacial cycles) on timescales far shorter than the slow forcing (orbital parameters, CO_2) that drives them. A standard modelling paradigm represents a climate variable X_t (e.g. thermohaline circulation strength, or global mean temperature anomaly) by a double-well potential $U_s(x)$ depending on a slow control parameter s (e.g. freshwater flux), with noise ϵ representing fast weather-scale fluctuations superimposed on the slow climate dynamics. Two mechanisms can produce a transition: *bifurcation-induced* tipping (the well at $x^\star$ disappears as s varies, a saddle-node bifurcation – a deterministic, zero-noise phenomenon) and *noise-induced* tipping (the well persists, $V^\star = 2\Delta U > 0$ remains positive, but $\mathbb{E}[\tau^\epsilon] \sim e^{V^\star/\epsilon}$ is finite and can be exceeded within the observation window even though no bifurcation occurs). Distinguishing these mechanisms from data – and estimating $V^\star$ from a limited observational record – is a major applied large-deviations problem, directly informing assessments of how close a system is to an irreversible "tipping."

Example — Worked example: early-warning signals via $V^\star(s)$

Suppose $U_s(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 - sx$ (a double well tilted by control parameter s). For small $|s|$, the two wells persist near $x_\pm(s) \approx \pm 1 + \frac{s}{2}$ with barrier height $\Delta U(s) \approx \frac{1}{4} - s$

(to first order in s, via perturbation of the unperturbed $s = 0$ values from Section 18.1). The quasi-potential from the right well to the barrier, $V_+^\star(s) = 2\Delta U(s) \approx \frac{1}{2} - 2s$, *decreases linearly* as s increases towards the bifurcation value s_c (where the right well disappears entirely, $\Delta U \to 0$). Mean escape time $\mathbb{E}[\tau^\epsilon] \sim e^{V_+^\star(s)/\epsilon}$ thus decreases super-exponentially in s as $s \to s_c^-$ even before the deterministic bifurcation – this is the theoretical basis for "early warning signals" (increasing variance, increasing autocorrelation, increased frequency of large excursions) observed in time series approaching tipping points: as $V^\star(s) \to 0$, noise-induced excursions over the (still-existing but shrinking) barrier become exponentially more frequent.

Genetic switches and cell differentiation. Gene regulatory networks with mutual repression (e.g. the classic toggle switch) exhibit bistability: two stable expression states (genes A "on"/B "off" and vice versa) separated by an unstable intermediate state, with stochasticity arising from the discreteness of molecule numbers (intrinsic noise, $\epsilon \sim 1/\Omega$ where Ω is the system volume/molecule-count scale). The Freidlin–Wentzell quasi-potential $V(x)$ over the concentration space gives the rate of spontaneous switching between cell fates – a phenomenon with direct relevance to stem-cell differentiation and to the reliability of synthetic biology circuits, where unwanted spontaneous switching is a "failure mode" in the Volume II sense.

Engineering: buckling and structural metastability. Slender structures under compressive load can have multiple stable equilibrium configurations (e.g. a shallow arch that can snap between "up" and "down" configurations). Ambient vibration (noise) can induce spontaneous snap-through even below the deterministic buckling load – a noise-induced transition whose rate is governed by the quasi-potential of the structural energy landscape, directly analogous to the double-well examples above, with ϵ representing the (small) vibrational energy scale relative to the elastic energy barrier. This connects the abstract machinery of this Part directly to the structural-reliability themes of Volume II.

Synthesis. Across all these examples, the same template recurs: (1) identify the deterministic dynamics $b = -\nabla U$ (or general b) and its metastable states; (2) compute the quasi-potential V (via the gradient formula $2\Delta U$ when applicable, or the Hamilton–Jacobi equation otherwise); (3) the rate of rare, large, qualitative transitions is $e^{-V^\star/\epsilon}$, with ϵ the relevant small-noise parameter of the application (thermal energy, $1/\Omega$, vibration amplitude, ...). Part 19 now turns to the computational side: since $e^{-V^\star/\epsilon}$ can be astronomically small, *direct* Monte Carlo simulation of such transitions is hopeless, and dedicated rare-event simulation algorithms – importance sampling, splitting, and adaptive methods, all exploiting the structure of $I_{[0,T]}$ and V revealed by this Part – are required.

PART 19

Rare-Event Simulation

Computational techniques for estimating probabilities of events too rare for naive Monte Carlo.

EXAMPLE — Exercises

1. For $U_s(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 - sx$ with $s = 0.1$, estimate $\Delta U(s) \approx \frac{1}{4} - s = 0.15$ and $V_+^\star(s) \approx 2 \times 0.15 = 0.3$. Compare $\mathbb{E}[\tau^\epsilon]$ at $\epsilon = 0.05$ for $s = 0$ vs. $s = 0.1$. *Solution:* At $s = 0$: $V^\star = 0.5$, $e^{0.5/0.05} = e^{10} \approx 22026$. At $s = 0.1$: $V^\star \approx 0.3$, $e^{0.3/0.05} = e^6 \approx 403.4$ – a $\sim 55\times$ reduction in mean escape time from a modest $s = 0.1$ shift, illustrating how sensitive noise-induced transition rates are to slow changes in a control parameter,

even far from the bifurcation point s_c.
2. Using the linear approximation $\Delta U(s) \approx 0.25 - s$, $\Delta U = 0$ at $s_c = 0.25$. Compute $V_+^\star(s)$ at $s = 0.2$ (close to s_c) and the corresponding escape time at $\epsilon = 0.05$. *Solution:* $\Delta U(0.2) \approx 0.25 - 0.2 = 0.05$, $V_+^\star(0.2) \approx 0.1$, $e^{0.1/0.05} = e^2 \approx 7.39$ – mean escape time $\approx \pi\sqrt{2} \times 7.39 \approx 32.8$, i.e. transitions that take ~ 22000 time units at $s = 0$ now take only ~ 33 time units at $s = 0.2$ – a $\sim 670\times$ acceleration, the "early warning" signature of approaching s_c.
3. Discuss qualitatively what an observer monitoring X_t would see as $s \to s_c$, in terms of the frequency of large excursions. *Solution:* As $V_+^\star(s) \to 0$, $\mathbb{E}[\tau^\epsilon] \to$ a finite, $O(1)$ value (the prefactor remains finite) – the system begins to make frequent excursions over the (vanishing) barrier well *before* the barrier disappears entirely. An observer would see increasingly frequent "flickering" between the two wells, increasing variance, and increasing autocorrelation – the classical early-warning signals, all directly explainable via $V^\star(s) \to 0$.

Further perspectives. The distinction between bifurcation-induced and noise-induced tipping, central to this section's climate application, has direct analogues in financial systems: a "bifurcation-induced" market crash might correspond to a fundamental change in market structure (e.g. a regulatory change removing a stabilising mechanism) after which the previous equilibrium ceases to exist, while a "noise-induced" crash occurs via a large but not-impossible sequence of adverse price movements crossing a threshold (e.g. triggering margin calls and forced liquidations) even though the pre-crash equilibrium remains technically stable. This financial reading connects directly to Volume IV's cascading-failure framework: the "noise" driving a noise-induced financial tipping event is often itself the result of a cascade (one institution's distress causing price moves that stress others), so the small-noise parameter ϵ in a financial application might represent not exogenous randomness but the magnitude of an initial, otherwise-containable shock – the quasi-potential $V^\star$ then quantifies the "shock-absorption capacity" of the financial system, with $\mathbb{E}[\tau^\epsilon] \sim e^{V^\star/\epsilon}$ interpreted as "how large must an initial shock be (relative to the system's absorption capacity, $1/\epsilon^{-1} \sim$ shock-to-capacity ratio) before cascading failure becomes likely within a relevant time horizon" – a direct bridge between the abstract noise-induced-transition machinery of this Part and the network-resilience themes that Volume IV develops in full.

Example — Application note: early-warning monitoring for a power grid

Power grids exhibit bistability under high load: a "normal operation" state and a "cascading blackout" state, separated by a stability margin that shrinks as load approaches transmission capacity (analogous to the control parameter s in Section 18.5's tilted double-well example). Grid operators monitor "voltage stability margins" $V^\star(s)$ (here s = current load relative to capacity) computed from power-flow models; as s increases toward a critical loading s_c, $V^\star(s) \to 0$, and – exactly as in the climate early-warning discussion – the frequency of small "near-miss" voltage excursions (transient dips that recover) should increase *before* an actual cascading blackout occurs, providing a real-time early-warning signal. The quantitative relationship $\mathbb{E}[\tau^\epsilon] \sim Ce^{V^\star(s)/\epsilon}$ (with ϵ representing the magnitude of routine load fluctuations, e.g. from renewable generation variability) means that monitoring the *empirical frequency* of near-miss excursions over a sliding time window provides an indirect estimate of $V^\star(s)/\epsilon$ – via $1/\widehat{\mathbb{E}[\tau^\epsilon]} \approx C^{-1}e^{-\hat{V}^\star(s)/\epsilon}$, so $\hat{V}^\star(s)/\epsilon \approx -\log(1/\widehat{\mathbb{E}[\tau^\epsilon]}) + \log C$ – and a sharp recent increase in near-miss frequency translates, via this logarithmic relationship, into a quantitative estimate of how much $V^\star(s)$ has decreased, i.e. how much "stability margin"

has been eroded by recent load growth or generation changes. Renewable-heavy grids, with their increased ϵ (fluctuating wind/solar output increases the effective noise intensity relative to older, more predictable thermal-generation-dominated grids), are – for a *fixed* $V^\star(s)$ – operating at effectively smaller $V^\star(s)/\epsilon$, i.e. closer (in the relevant exponential sense) to instability, even without any change in the underlying deterministic stability margin $V^\star(s)$ itself – a direct, quantitative articulation of the qualitative concern that "increased variability requires increased margins," translating an engineering intuition into the precise ϵ-dependence of the Freidlin–Wentzell exit-time formula.

Algorithm: early-warning indicator computation from time-series data

Input: time series $\{X_t\}_{t=1}^{T}$ from a system suspected to be approaching a noise-induced or bifurcation-induced transition.

1. Detrend the series (remove any slow deterministic trend, e.g. via a moving-average or spline fit), retaining the "fluctuation" residual $\tilde{X}_t = X_t - \text{trend}(t)$.
2. Compute, in a sliding window of length w (e.g. $w = T/10$): (a) the variance $\text{Var}(\tilde{X}_t)$ within the window; (b) the lag-1 autocorrelation $\rho_1 = \text{Corr}(\tilde{X}_t, \tilde{X}_{t+1})$ within the window.
3. Plot $\text{Var}(\tilde{X}_t)$ and ρ_1 as functions of the window's centre time; a statistically significant *increasing trend* in either (assessed e.g. via Kendall's τ rank correlation between the windowed statistic and time, with a permutation-test p-value) is the standard "early warning signal."
4. Relate to $V^\star(s)$: under the linearised dynamics near a well, $\text{Var}(\tilde{X}) \approx \epsilon/(2|U_s''(x^\star)|)$ and $\rho_1 \approx e^{-|U_s''(x^\star)|\Delta t}$ (an Ornstein–Uhlenbeck-type approximation), so $U_s''(x^\star) \to 0$ (the well flattening as $s \to s_c$) drives both $\text{Var} \to \infty$ and $\rho_1 \to 1$ – the two most commonly used early-warning indicators are both, in this sense, direct (if indirect, via the linearisation) probes of the same underlying quantity, the curvature $U_s''(x^\star)$ that also enters the Eyring–Kramers prefactor (Section 18.4).

A well-documented pitfall: trends in Var and ρ_1 can arise from causes *unrelated* to an approaching tipping point (e.g. a genuine, gradual change in the noise intensity ϵ itself, with $U_s''(x^\star)$ constant) – the indicators are necessary but not sufficient signatures of $V^\star(s) \to 0$, and should be interpreted alongside independent information about whether s (the control parameter) is itself changing, and in which direction, lest a change in noise characteristics be misattributed to an approaching bifurcation.

Key formulas: Section 18.5

- Tilted potential: $U_s(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 - sx$; $\Delta U(s) \approx \frac{1}{4} - s$ (linear approx.), $s_c = \frac{1}{4}$.
- $V_+^\star(s) \approx 2\Delta U(s) = \frac{1}{2} - 2s$; $\mathbb{E}[\tau^\epsilon] \sim e^{V_+^\star(s)/\epsilon}$ decreases super-exponentially as $s \to s_c^-$.
- Early-warning signals (Section 18.5): $\text{Var} \propto 1/|U_s''(x^\star)|$, $\rho_1 \approx e^{-|U_s''(x^\star)|\Delta t}$; both diverge/saturate as $U_s''(x^\star) \to 0$.

Notation and further reading. "Early warning signals" (variance, autocorrelation) are sometimes collectively called "critical slowing down" indicators in the dynamical-systems and ecology literature, reflecting that both signals arise from the same underlying slowing of the system's relaxation rate ($|U_s''(x^\star)| \to 0$) as a bifurcation is approached. Scheffer et

al.'s widely-cited 2009 *Nature* review "Early-warning signals for critical transitions" surveys applications across ecology, climate, and (briefly) finance; for the underlying stochastic-dynamics theory, Kuehn's *A Mathematical Framework for Critical Transitions* connects the bifurcation-theoretic and stochastic perspectives rigorously.

Cross-volume connections. The bifurcation-induced vs. noise-induced tipping distinction, and the early-warning-signal machinery built on it, directly parallels Volume IV's treatment of cascading failures: a "bifurcation-induced cascade" corresponds to a network reaching a structural threshold (e.g. losing connectivity) where cascade becomes inevitable regardless of further perturbation, while a "noise-induced cascade" occurs via an unlucky sequence of individually-survivable shocks crossing a (still nominally adequate) capacity threshold – the same early-warning indicators (increasing "variance" and "autocorrelation" of network stress metrics) are, in fact, used in power-grid and financial-network monitoring, as this section's application note discusses. Volume VI's antifragility framework can be read as asking: "how can a system's $V^\star(s)$ be *increased* (made more robust) as a function of design choices, rather than merely monitored as s drifts toward s_c?" – shifting from the monitoring perspective of this section to a design perspective.

Example — Extended discussion: hysteresis and forward/backward critical points

For $U_s(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2 - sx$, the deterministic double-well structure persists for $|s| < s_c = \frac{2}{3\sqrt{3}} \approx 0.385$ (found from $U_s'(x) = 0$ and $U_s''(x) = 0$ simultaneously, $x = \pm 1/\sqrt{3}$).
Hysteresis loop. Starting at $s = 0$ in the left well and increasing s: the left well persists until $s = +s_c \approx 0.385$, where the system jumps to the right well. Decreasing s from above s_c: the system remains in the right well until $s = -s_c \approx -0.385$, where it jumps back – a hysteresis loop of width $2s_c \approx 0.77$.
Interaction with noise-induced tipping. For $\epsilon > 0$, noise-induced jumps (Section 18.5's main mechanism) occur *before* s_c, at some $s^\star(\epsilon) < s_c$ with $s^\star(\epsilon) \to s_c$ as $\epsilon \to 0$. Noisy hysteresis loops are thus systematically *narrower* than the deterministic prediction. Distinguishing "narrow loop because of strong noise" from "intrinsically weak hysteresis (small s_c)" is a genuine inferential challenge – another instance of the attribution difficulty that motivates Section 20.4's model-uncertainty framework.

Example — Extended discussion: stochastic resonance

If the double-well potential of this section's hysteresis discussion is itself slowly, periodically time-varying ($s(t) = A\cos(\omega t)$, ω small relative to the noise-induced hopping rate), noise can *synchronise* with the forcing: hopping becomes most regular – closest to period $2\pi/\omega$ – at an intermediate, non-zero noise level, a phenomenon called stochastic resonance. Too little noise: hops too rare to track the forcing; too much: hops occur essentially at random, swamping the signal.
Risk interpretation. If $s(t)$ represents a slowly oscillating macroeconomic driver and the two wells represent "normal" and "stressed" regimes, stochastic resonance suggests regime-switching can appear most business-cycle-synchronised at *intermediate* volatility – neither very calm nor very turbulent markets – a non-monotonic relationship that purely deterministic or purely diffusive models would both miss.

19.1 Importance Sampling: Theory and Variance Reduction

Overview

Direct ("crude") Monte Carlo estimates $\Pr(X \in A)$ for rare A by counting hits in N i.i.d. samples – but if $\Pr(A) = 10^{-6}$, one needs $N \gtrsim 10^8$ samples for a handful of hits and a usable relative error. *Importance sampling* (IS) instead samples from a different ("proposal") distribution under which A is no longer rare, and corrects via a likelihood-ratio weight – in principle reducing the required sample size by orders of magnitude.

Definition: Importance sampling estimator

To estimate $p = \Pr_\nu(X \in A) = \mathbb{E}_\nu[\mathbf{1}_A(X)]$, choose a proposal density q (with $q(x) > 0$ wherever $\nu(x)\mathbf{1}_A(x) > 0$) and write

$$p = \int \mathbf{1}_A(x)\,\nu(x)\,dx = \int \mathbf{1}_A(x)\,\frac{\nu(x)}{q(x)}\,q(x)\,dx = \mathbb{E}_q\Big[\mathbf{1}_A(X)\,L(X)\Big], \qquad L(x) = \frac{\nu(x)}{q(x)} \text{ (likelihood ratio).}$$

The IS estimator is $\hat{p}_N = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_A(X_i)L(X_i)$, $X_i \overset{\text{iid}}{\sim} q$ – unbiased for any valid q ($\mathbb{E}_q[\hat{p}_N] = p$), but its *variance* depends critically on the choice of q.

Example — Theorem: Variance of the IS estimator and the zero-variance proposal

$$\mathrm{Var}_q(\hat{p}_N) = \frac{1}{N}\Big(\mathbb{E}_q[\mathbf{1}_A(X)L(X)^2] - p^2\Big) = \frac{1}{N}\Big(\int_A \nu(x)L(x)\,dx - p^2\Big).$$

The *zero-variance* (ideal) proposal is the *conditional* distribution $q^\star(x) = \nu(x)\mathbf{1}_A(x)/p$ (i.e. ν restricted and renormalised to A): then $L^\star(x) = \nu(x)/q^\star(x) = p$ for $x \in A$, a constant, so $\hat{p}_N \equiv p$ for every sample – exactly zero variance. Of course $q^\star$ requires knowing p (circular!), but it identifies the *target*: a good proposal should approximate $\nu(\cdot \mid A)$, concentrating mass where A "lives" rather than where ν's bulk lives.

Relative error and the bounded relative error (BRE) property. Define the relative error $\mathrm{RE} = \sqrt{\mathrm{Var}(\hat{p}_N)}/(p\sqrt{N})$ – the natural scale-free measure of estimator quality for small p. Crude Monte Carlo has $\mathrm{RE} \to \infty$ as $p \to 0$ (since $\mathrm{Var} = \frac{1}{N}p(1-p) \approx \frac{p}{N}$, giving $\mathrm{RE} \approx 1/\sqrt{Np} \to \infty$). A "good" IS scheme achieves *bounded relative error*: RE remains bounded as $p \to 0$ (e.g. as the rarity parameter, large deviation scale $1/\epsilon$ or n, $\to \infty$). Designing BRE estimators is the central technical challenge of rare-event simulation, and large deviation theory – specifically, the exponential tilting of Section 17.2/19.2 – is the principal tool for constructing them.

Example — Worked example: importance sampling for a Gaussian tail

Let $X \sim \mathcal{N}(0,1)$, $A = \{X > 4\}$, $p = \Pr(X > 4) \approx 3.17 \times 10^{-5}$ (so crude MC needs $N \sim 10^7$ for a few hits). Choose proposal $q = \mathcal{N}(\theta, 1)$ for some shift θ (a "mean shift" – the simplest non-trivial IS scheme). The likelihood ratio is $L(x) = \frac{\phi(x)}{\phi(x-\theta)} = e^{-\theta x + \theta^2/2}$ (where ϕ is the standard normal density). Variance:

$$\mathbb{E}_q[\mathbf{1}_A(X)L(X)^2] = \int_4^\infty \phi(x-\theta)\,e^{-2\theta x+\theta^2}\,dx = e^{-\theta^2}\int_4^\infty \frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2-2\theta x+2\theta^2+\theta^2-(x-\theta)^2/2+(x-\theta)^2/2}\ldots$$

A cleaner route: $\phi(x-\theta)e^{-2\theta x+\theta^2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2-2\theta x+\theta^2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2+\theta x-\frac{1}{2}\theta^2-2\theta x+\theta^2} =$

$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2-\theta x+\frac{1}{2}\theta^2} = \phi(x)e^{-\theta x+\theta^2/2} = \phi(x+\theta)e^{\theta^2}$ (completing the square again, shifting by $-\theta$). So $\mathbb{E}_q[\mathbf{1}_A L^2] = e^{\theta^2}\Pr(X > 4-\theta)$ for $X \sim \mathcal{N}(0,1)$.
Choosing $\theta = 4$ (shifting the proposal mean to land exactly at the threshold) gives $\mathbb{E}_q[\mathbf{1}_A L^2] = e^{16}\Pr(X > 0) = e^{16}/2 \approx 4.4 \times 10^6$. Then $\mathrm{Var}(\widehat{p}_N) \approx \frac{1}{N}(4.4 \times 10^6 - (3.17 \times 10^{-5})^2) \approx \frac{4.4\times 10^6}{N}$, vs. crude MC's Var $\approx \frac{p}{N} = \frac{3.17\times 10^{-5}}{N}$ – IS variance is *larger*?! The resolution: compare relative errors. $\mathrm{RE}_{\mathrm{IS}} = \sqrt{4.4 \times 10^6}/(3.17 \times 10^{-5}\sqrt{N}) \approx 2097/(3.17 \times 10^{-5}\sqrt{N}) \approx 6.6 \times 10^7/\sqrt{N}$, while $\mathrm{RE}_{\mathrm{crude}} = \sqrt{p}/(p\sqrt{N}) = 1/\sqrt{pN} \approx 1/\sqrt{3.17 \times 10^{-5} N} \approx 178/\sqrt{N}$. So crude MC actually has *smaller* relative error here?! This reflects that for a *scalar* Gaussian tail, the naive shift $\theta = 4$ is not yet optimal – Section 19.2 shows that the exponentially-tilted (large-deviation-optimal) shift, here $\theta^\star = 4$ exactly by Cramér's theorem ($\Lambda^{\star\prime}(4) = 4$ for $\mathcal{N}(0,1)$ since $\Lambda(\lambda) = \lambda^2/2$, $\Lambda'(\lambda) = \lambda$), *does* achieve bounded relative error in the correct asymptotic regime (as the threshold $\to \infty$, not fixed at 4) – the apparent paradox above arises because $\theta = 4$ happens to be exactly the LDP-optimal tilt, and the residual $O(1)$ prefactor (not captured by the exponential rate alone) is what the variance calculation is sensitive to at this finite threshold. The qualitative lesson stands: *naive* shifts (e.g. $\theta = 2$, undershooting) perform far worse, with variance ratios growing without bound as the threshold increases.

Outlook. The example illustrates both the promise and the subtlety of importance sampling: the right exponential tilt (Section 19.2) is asymptotically optimal and provably achieves bounded relative error as rarity $\to \infty$, but getting the tilt parameter right matters, and for high-dimensional or path-space problems (Sections 19.3–19.5) identifying the optimal tilt analytically is often infeasible, motivating adaptive and simulation-based alternatives.

Example — Exercises

1. For a discrete example, $\nu = \mathrm{Bernoulli}(0.01)$ (rare $X = 1$ with $p = 0.01$), and $A = \{X = 1\}$. Write down the zero-variance proposal $q^\star$. *Solution:* $q^\star(x) = \nu(x)\mathbf{1}_A(x)/p$. Since $A = \{1\}$ and $p = \nu(1) = 0.01$, $q^\star(1) = \nu(1) \times 1/0.01 = 1$ and $q^\star(0) = 0$ – i.e. $q^\star$ is the point mass at $X = 1$: the "proposal" simply always returns the rare outcome, with weight $L^\star = \nu(1)/q^\star(1) = 0.01/1 = 0.01 = p$ exactly, as the theory predicts.
2. For crude MC with $p = 10^{-4}$ and $N = 10^4$, compute $\mathrm{RE}_{\mathrm{crude}} = 1/\sqrt{Np}$. *Solution:* $\mathrm{RE} = 1/\sqrt{10^4 \times 10^{-4}} = 1/\sqrt{1} = 1$ – a relative error of 100%, i.e. the estimate's standard deviation is comparable to its own value – essentially useless. To reach $\mathrm{RE} = 0.1$ for the same p would require $N = 10^6$.
3. For $X \sim \mathcal{N}(0,1)$, verify the likelihood ratio identity $\phi(x-\theta)e^{-2\theta x+\theta^2} = \phi(x+\theta)e^{\theta^2}$ used in the worked example, by direct substitution for $\theta = 1, x = 2$. *Solution:* LHS: $\phi(1)e^{-4+1} = \phi(1)e^{-3}$. RHS: $\phi(3)e^1$. Numerically, $\phi(1) \approx 0.2420$, so LHS $\approx 0.2420 \times e^{-3} \approx 0.2420 \times 0.0498 \approx 0.01205$. $\phi(3) \approx 0.00443$, so RHS$\approx 0.00443 \times e \approx 0.00443 \times 2.718 \approx 0.01205$ – equal, confirming the identity.

Further perspectives. The "bounded relative error" (BRE) property, while desirable, is not always achievable: for some rare-event problems (particularly those involving heavy-tailed, subexponential distributions – precisely the regime that falls outside Cramér/Gärtner–Ellis, as noted in Section 17.4's remarks), *no* exponential-family change of measure achieves BRE, and specialised techniques (e.g. conditional Monte Carlo, conditioning on the index of the "big jump" in a sum of heavy-tailed variables, exploiting the one-big-jump principle of Section 20.5 directly as an algorithmic device) are needed instead. This is an important caveat to the otherwise rosy picture of importance sampling: the exponential-tilting machinery

of Section 19.2 is powerful precisely in the "light-tailed" (Cramér) regime that this Part's large-deviation theory addresses, but tail-risk applications (Volume I) frequently involve heavy-tailed losses for which a different algorithmic toolkit (still informed by asymptotic theory, but of the "regular variation" rather than "large deviation" flavour) is required. The variance formula $\mathrm{Var}_q(\hat{p}_N) = \frac{1}{N}(\mathbb{E}_q[\mathbf{1}_A L^2] - p^2)$ itself is completely general and remains the correct starting point for analysing *any* importance sampling scheme, light- or heavy-tailed; what differs is which proposals q make the second moment $\mathbb{E}_q[\mathbf{1}_A L^2]$ controllable.

Example — Application note: variance budget for a Monte Carlo risk engine

A risk management system computes daily VaR via Monte Carlo simulation of $N = 10^5$ scenarios from a portfolio risk model, targeting the 99.5% quantile ($p = 0.005$). The relative error of the empirical quantile estimator is, for a quantile q, approximately $\mathrm{RE} \approx \sqrt{q(1-q)}/(f(x_q)\sqrt{N}\,x_q)$ where f is the density at the quantile – a more complex formula than the simple tail-probability $\mathrm{RE} = 1/\sqrt{Np}$ of Section 19.1, but with the same qualitative $1/\sqrt{N}$ scaling and the same qualitative degradation as $q \to 1$ (via $f(x_q) \to 0$ for light-tailed distributions, and $x_q \to \infty$). For a more extreme regulatory quantile, e.g. 99.98% ($p = 0.0002$, a "1-in-5000-day," roughly 1-in-20-year, event), the same $N = 10^5$ scenarios give an effective "sample size near the quantile" of only $N \times p = 20$ – comparable to the crude-MC "few hits" problem of Section 19.1's opening discussion, even though $N = 10^5$ sounds large. A risk engine targeting reliable 99.98% quantile estimates would need either $N \gtrsim 10^7$ (to get ~ 2000 effective samples near the quantile, a reasonable target for a $\sim 2\%$ relative error on the quantile itself) or an importance-sampling scheme (Section 19.2) tilting the scenario generator toward the tail – a direct, practical illustration of why regulatory capital models for extreme quantiles (e.g. Basel's move toward Expected Shortfall at 97.5%, or insurance Solvency II's 99.5% VaR) increasingly mandate or strongly encourage variance-reduction techniques, since "just run more crude scenarios" becomes computationally prohibitive for the most extreme (and most capital-relevant) quantiles.

Algorithm: generic importance sampling estimator with diagnostics

Input: target probability $p = \Pr_\nu(X \in A)$; proposal q; sample size N.

1. Draw $X_1, \dots, X_N \overset{\text{iid}}{\sim} q$.
2. Compute weights $L_i = \nu(X_i)/q(X_i)$ and indicators $\mathbf{1}_A(X_i)$.
3. $\hat{p} = \frac{1}{N}\sum_i L_i \mathbf{1}_A(X_i)$; $\widehat{\mathrm{Var}}(\hat{p}) = \frac{1}{N(N-1)}\sum_i (L_i \mathbf{1}_A(X_i) - \hat{p})^2$.
4. **Diagnostic (effective sample size):** compute $\mathrm{ESS} = \frac{(\sum_i L_i \mathbf{1}_A(X_i))^2}{\sum_i (L_i \mathbf{1}_A(X_i))^2}$ (or, more generally, $\mathrm{ESS} = (\sum L_i)^2 / \sum L_i^2$ over all samples, not just those in A). $\mathrm{ESS} \ll N$ signals that a small number of samples with very large weights dominate the estimate – a red flag for high variance not fully captured by the naive variance formula (which itself has high variance when weights are heavy-tailed).

The ESS diagnostic in step 4 is essential precisely because $\mathrm{Var}(\hat{p})$ itself, when weights L_i are heavy-tailed (e.g. if the proposal q is poorly chosen, with $q(x) \to 0$ faster than $\nu(x)$ in some region, making $L = \nu/q \to \infty$ there), can have infinite *true* variance even though the sample variance $\widehat{\mathrm{Var}}(\hat{p})$ computed from any finite sample is itself finite – a classic "estimating the variance of an estimator whose variance may not exist" pathology. A small ESS (say, $\mathrm{ESS} < N/10$) is a practical signal to revisit the proposal q (e.g. via the exponential tilting of Section 19.2 or the adaptive methods of Sections 19.4–19.5) rather than simply increasing N, which would not resolve an underlying heavy-tailed-weight problem.

Key formulas: Section 19.1

- IS identity: $p = \mathbb{E}_q[\mathbf{1}_A(X)L(X)]$, $L = \nu/q$, $\hat{p}_N = \frac{1}{N}\sum_i \mathbf{1}_A(X_i)L(X_i)$.
- Zero-variance proposal: $q^\star(x) = \nu(x)\mathbf{1}_A(x)/p$, giving $L^\star \equiv p$ on A.
- $\mathrm{Var}_q(\hat{p}_N) = \frac{1}{N}(\mathbb{E}_q[\mathbf{1}_A L^2] - p^2)$.
- Relative error: $\mathrm{RE} = \sqrt{\mathrm{Var}(\hat{p}_N)}/(p\sqrt{N})$; crude MC RE $\approx 1/\sqrt{Np} \to \infty$ as $p \to 0$.

Notation and further reading. "Importance sampling" is sometimes abbreviated IS throughout the simulation literature (as in this volume from Section 19.2 onward); the "proposal" distribution q is also called the "importance distribution," "biasing distribution," or (in sequential Monte Carlo contexts) the "instrumental distribution." Asmussen & Glynn's *Stochastic Simulation: Algorithms and Analysis* provides a comprehensive treatment of variance reduction techniques including importance sampling, with a particular focus on the rare-event setting central to this Part.

Cross-volume connections. The "bounded relative error" criterion for rare-event estimators is the simulation-side analogue of the "well-posedness" criteria Volume III applies to robust optimisation formulations: just as a robust optimisation problem should have a well-defined (finite, interpretable) worst-case value as the ambiguity set shrinks to a point, a rare-event estimator should have bounded relative error as the target probability shrinks to zero – both are "the answer shouldn't degenerate as the regime becomes more extreme" requirements, reflecting a shared concern (across decision-theoretic and computational contexts) for methods that remain meaningful in genuinely extreme regimes, not just moderate ones.

Example — Extended discussion: stratified sampling as a complementary technique

Stratified sampling partitions the sample space into strata $\{S_1, \dots, S_K\}$ with known $w_k = \Pr(S_k)$, draws N_k samples from $\nu(\cdot|S_k)$, and estimates $\Pr(A) = \sum_k w_k \hat{p}_k$. With proportional allocation $N_k = Nw_k$, $\mathrm{Var}(\hat{p}) = \frac{1}{N}\sum_k w_k \Pr(A|S_k)(1 - \Pr(A|S_k)) \le \frac{1}{N}\Pr(A)(1-\Pr(A))$ – never worse than crude MC, strictly better when $\Pr(A|S_k)$ varies across strata.

Worked example. For Appendix G's desk, stratify by "fraction of stressed days" into low/typical/high strata ($w \approx 0.3/0.5/0.2$) with very different $\Pr(A|S_k) \approx 0.01/0.10/0.40$ for $A = \{S_{250} \le -5\}$. Crude MC's variance is dominated by this huge between-stratum spread; proportional-allocation stratification eliminates it entirely, retaining only the smaller within-stratum binomial variances.

Combining with IS. If a more extreme $A' = \{S_{250} \le -15\}$ has small $\Pr(A'|S_3)$ even within the high-stress stratum, importance sampling *within* S_3 refines further – a "stratify, then tilt within strata" two-level structure standard in catastrophe modelling (Section 20.5), where strata correspond to "number of major hurricanes" and IS refines the rare multi-hurricane strata.

Example — Extended discussion: variance allocation across multiple target events

When a single budget N must estimate several rare probabilities $\Pr(A_1), \dots, \Pr(A_K)$ simultaneously (e.g. breach probabilities for K risk limits), no single q is simultaneously optimal for all A_k. A practical compromise is a mixture proposal $q = \sum_k w_k q_k^\star$, weights w_k chosen e.g. proportional to $1/\sqrt{\Pr(A_k)}$, allocating relatively more "attention" to rarer events.

Bookkeeping. Each path simulated under q carries a likelihood-ratio weight ν/q

usable for *all* K estimators simultaneously (multiple importance sampling), so the marginal cost of K probabilities together is far below $K\times$(cost of one alone) – directly relevant whenever a stress-testing exercise (Section 20.5) must report breach probabilities across a whole suite of limits, not one at a time.

19.2 Exponential Tilting and the Optimal Change of Measure

Overview

Exponential tilting – the same change of measure that appears in the proof of Cramér's theorem (Section 17.2) – provides a principled, large-deviation-optimal recipe for the importance sampling proposal q: tilt the original distribution ν by $e^{\lambda x}$ for the λ that makes the rare event A *typical* under the tilted law.

Definition: Exponentially tilted family

For ν with CGF $\Lambda(\lambda) = \log \mathbb{E}_\nu[e^{\lambda X}] < \infty$, the *exponential tilt* (Esscher transform) at parameter λ is the probability measure

$$\nu_\lambda(dx) = e^{\lambda x - \Lambda(\lambda)}\,\nu(dx).$$

This is a valid probability measure (integrates to 1 since $\Lambda(\lambda) = \log \int e^{\lambda x}\nu(dx)$), with mean $\mathbb{E}_{\nu_\lambda}[X] = \Lambda'(\lambda)$ and the likelihood ratio $\frac{d\nu}{d\nu_\lambda}(x) = e^{-\lambda x + \Lambda(\lambda)}$.

Example — Theorem: Asymptotic optimality of the LDP-tilted proposal

To estimate $p_n = \Pr(\bar{X}_n \geq x)$ for $x > \mu = \mathbb{E}[X]$, with $X_1, \dots, X_n$ i.i.d. $\sim \nu$, sample instead from $\nu_{\lambda^\star}^{\otimes n}$ where $\lambda^\star = \lambda^\star(x)$ solves $\Lambda'(\lambda^\star) = x$ (the LDP-optimal tilt of Section 17.2, making x the *mean* under $\nu_{\lambda^\star}$, hence $\{\bar{X}_n \geq x\}$ typical). The likelihood-ratio weight for a sample path $(X_1, \dots, X_n)$ is

$$L = \prod_{i=1}^{n} \frac{d\nu}{d\nu_{\lambda^\star}}(X_i) = e^{-\lambda^\star \sum_i X_i + n\Lambda(\lambda^\star)} = e^{-n(\lambda^\star \bar{X}_n - \Lambda(\lambda^\star))}.$$

On the event $\{\bar{X}_n \geq x\}$ (now typical, $\bar{X}_n \approx x$ under $\nu_{\lambda^\star}$), $L \approx e^{-n(\lambda^\star x - \Lambda(\lambda^\star))} = e^{-n\Lambda^*(x)} = e^{-nI(x)} \approx p_n$ – the weight itself is $\approx p_n$ for every "typical" sample, exactly the zero-variance behaviour identified in Section 19.1 ($L^\star = p$ constant on A). This is the sense in which exponential tilting is *asymptotically* optimal: as $n \to \infty$, the relative error of the tilted IS estimator stays bounded, whereas crude MC's relative error $\to \infty$.

Example — Worked example: tilted estimation of a sum-exceedance probability

Let $X_i \sim \text{Exp}(1)$ i.i.d., and estimate $p_n = \Pr(\bar{X}_n \geq 3)$ for $n = 20$ (so $S_n = \sum X_i \geq 60$, a $3\times$-mean exceedance for a sum with mean 20). From Section 17.2, $\Lambda(\lambda) = -\log(1-\lambda)$ for $\lambda < 1$, and $\Lambda^*(x) = \theta x - 1 - \log(\theta x)$ with $\theta = 1$: $\Lambda^*(3) = 3 - 1 - \log 3 \approx 0.901$. The optimal tilt solves $\Lambda'(\lambda^\star) = 1/(1-\lambda^\star) = 3 \Rightarrow \lambda^\star = 2/3$. Under the tilt, $\nu_{\lambda^\star}(dx) = e^{(2/3)x - \Lambda(2/3)} e^{-x} dx$; since $\Lambda(2/3) = -\log(1/3) = \log 3$, $\nu_{\lambda^\star}(dx) = 3e^{(2/3)x} e^{-x} dx = 3e^{-x/3} dx$ – i.e. $\nu_{\lambda^\star} = \text{Exp}(1/3)$, an exponential with mean 3 (exactly $x = 3$, as the theorem predicts: $\Lambda'(\lambda^\star) = \mathbb{E}_{\nu_{\lambda^\star}}[X] = 3$). So: *simulate* $X_i \sim \text{Exp}(1/3)$ *instead of* $\text{Exp}(1)$ –

under this law, $\bar{X}_{20} \geq 3$ is now a typical (probability ≈ 0.5) event, and each sample's weight is $L = \prod_i \frac{d\text{Exp}(1)}{d\text{Exp}(1/3)}(X_i) = \prod_i \frac{e^{-X_i}}{3e^{-X_i/3}} = \frac{1}{3^{20}} e^{-\sum X_i(1-1/3)} = 3^{-20} e^{-\frac{2}{3}S_n}$. For a typical sample with $S_n \approx 60$, $L \approx 3^{-20}e^{-40} \approx 2.9 \times 10^{-10} \times 4.2 \times 10^{-18} \approx 1.2 \times 10^{-27}$ – absurdly tiny numbers, but each weighted indicator $\mathbf{1}_A(X)L(X)$ has expectation exactly $p_{20} = e^{-20 \times 0.901} \approx e^{-18.0} \approx 1.5 \times 10^{-8}$, and with $N = 10^4$ tilted samples (each landing in A with probability ≈ 0.5, so ~ 5000 "hits" contribute non-trivially), the IS estimate of p_{20} achieves a relative error far smaller than crude MC could achieve even with $N = 10^9$ (where the expected number of hits would be only ~ 15).

Multidimensional and path-space tilting. The recipe generalises directly: for $\mathbb{R}^d$-valued sums, tilt by $e^{\square \cdot \mathbf{x} - \Lambda(\square)}$ with $\square^\star$ solving $\nabla\Lambda(\square^\star) = \mathbf{x}$; for path-space rare events (Section 18.1), the analogous "tilt" is a *change of drift* $b \to b + au^\star$ where $u^\star$ is the control achieving the minimal-action path $\phi^\star$ (the Girsanov tilting alluded to in Section 18.1) – this is the basis of the subsolution approach to importance sampling for diffusions, where one constructs $u^\star$ from an approximate solution of the Hamilton–Jacobi equation for the quasi-potential V (Section 18.2). When Λ or V cannot be computed in closed form – the typical case in genuinely high-dimensional or non-gradient systems – Sections 19.3–19.5 provide simulation-based alternatives that do not require an explicit tilt.

Example — Exercises

1. For $X \sim \text{Gamma}(k, \theta)$ (shape k, rate θ), $\Lambda(\lambda) = -k\log(1 - \lambda/\theta)$ for $\lambda < \theta$. Find the tilted distribution ν_λ. *Solution:* $\nu_\lambda(dx) \propto e^{\lambda x}\nu(dx) \propto e^{\lambda x}x^{k-1}e^{-\theta x} = x^{k-1}e^{-(\theta-\lambda)x}$ – proportional to a $\text{Gamma}(k, \theta - \lambda)$ density, i.e. the tilt simply shifts the rate parameter from θ to $\theta - \lambda$, preserving the shape k.
2. Verify $\mathbb{E}_{\nu_\lambda}[X] = \Lambda'(\lambda)$ for the Gamma case. *Solution:* $\Lambda'(\lambda) = \frac{k}{\theta - \lambda}$. And $\mathbb{E}_{\text{Gamma}(k,\theta-\lambda)}[X] = k/(\theta - \lambda)$ (mean of a Gamma is shape/rate) – matching exactly.
3. For $\text{Exp}(1)$ ($\theta = 1$), find $\lambda^\star$ such that $\Lambda'(\lambda^\star) = 5$, and the corresponding tilted rate. *Solution:* $\Lambda'(\lambda) = 1/(1 - \lambda) = 5 \Rightarrow 1 - \lambda = 1/5 \Rightarrow \lambda^\star = 4/5 = 0.8$. Tilted rate $= \theta - \lambda^\star = 1 - 0.8 = 0.2$, i.e. $\nu_{\lambda^\star} = \text{Exp}(0.2)$, mean 5 – as required, the tilted mean equals the target $x = 5$.

Further perspectives. Exponential tilting (the Esscher transform) has an independent and much older history in actuarial mathematics, where Esscher (1932) introduced it as a technique for approximating aggregate claim distributions – decades before its large-deviation-optimality was understood. This actuarial origin is a nice illustration of a recurring pattern in applied mathematics: a technique discovered for one purpose (here, distributional approximation) is later found to be *exactly* the right tool for a seemingly unrelated purpose (variance-optimal importance sampling), once the underlying variational structure (here, the Legendre transform) is recognised as common to both. A further generalisation relevant to multi-step or path-dependent problems: the tilting parameter $\lambda^\star$ need not be constant – for path-space problems, the optimal "tilt" is generally a *time-varying* (state-dependent) drift change $u^\star(x, t)$, determined by the value function of an associated stochastic control problem (the Hamilton–Jacobi–Bellman equation for "minimum expected cost to reach the rare event"), of which the constant tilt $\lambda^\star$ for i.i.d. sums is the degenerate, state-independent special case. This stochastic-control perspective on importance sampling – "find the control that makes the rare event typical, at minimum L^2 cost" – unifies Sections 19.2–19.4 and is the modern framework within which adaptive importance sampling algorithms (including cross-entropy, Section 19.4) are increasingly analysed and designed.

Example — Application note: tilted scenario generation for stress testing

A bank's stress-testing framework generates macroeconomic scenarios (GDP growth, unemployment, interest rates) from a multivariate model with CGF $\Lambda(\square)$, and wishes to generate scenarios conditional on "severe recession" ($A = \{\text{GDP growth} \leq -3\%\}$, a roughly 1-in-50-year event under the baseline model, $p \approx 0.02$). Rather than running $N = 10^4$ baseline scenarios and retaining only the ~ 200 that satisfy A (a form of crude rejection sampling, wasteful but not hopeless at this moderate p), the bank applies the exponential tilt $\square^\star$ solving $\nabla\Lambda(\square^\star) = \mathbf{x}^\star$ where $\mathbf{x}^\star$ is a representative point in A (e.g. GDP growth $= -4\%$, with other variables at their tilted-conditional means). Generating $N = 10^4$ scenarios from the tilted model $\nu_{\square^\star}$, *all* of which are now representative of "severe recession" conditions (by construction, A is typical under $\nu_{\square^\star}$), each weighted by $L = e^{-\square^\star \cdot \mathbf{X} + \Lambda(\square^\star)}$, the bank obtains 10^4 "severe recession scenarios" with associated probability weights – useful not just for estimating $\Pr(A)$ itself (which was already roughly known, $p \approx 0.02$), but for estimating *conditional* quantities given A, e.g. $\mathbb{E}[\text{portfolio loss} \mid A]$, with the *full* 10^4-scenario resolution rather than the ~ 200 scenarios that crude rejection sampling would provide. This "tilt to make the conditioning event typical, then analyse the tilted ensemble with importance weights" pattern is the practical essence of exponential tilting for stress testing – it is not primarily about estimating $\Pr(A)$ more efficiently (often $\Pr(A)$ is reasonably well-known or even a modelling *input*, as in regulatory stress scenarios defined by a target probability), but about obtaining a richly-resolved *conditional* scenario ensemble for downstream analysis.

Algorithm: exponential tilting for sums of i.i.d. random variables

Input: i.i.d. summands $X_i \sim \nu$ with CGF Λ; target $x > \mathbb{E}[X]$; sample size N; number of summands n.

1. Compute $\lambda^\star$ solving $\Lambda'(\lambda^\star) = x$ via Algorithm 17.2's Newton iteration.
2. For each of N replicates: simulate $X_1, \dots, X_n \overset{\text{iid}}{\sim} \nu_{\lambda^\star}$ (the tilted distribution; for common families – Gaussian, Exponential, Gamma, Poisson – $\nu_{\lambda^\star}$ is in the *same family* with shifted parameters, per Section 19.2's exercises, so simulation is no harder than from ν itself).
3. Compute $S_n = \sum_i X_i$, $L = e^{-\lambda^\star S_n + n\Lambda(\lambda^\star)}$, and the indicator $\mathbf{1}_{\{S_n/n \geq x\}}$.
4. $\hat{p} = \frac{1}{N}\sum L\,\mathbf{1}$; report $\hat{p}$ along with the ESS diagnostic (Algorithm 19.1).

For exponential-family ν (covering most distributions used in practice – Gaussian, Exponential, Gamma, Poisson, Binomial, and others – all of which are exponential families and hence closed under exponential tilting), step 2 requires *no* new simulation machinery: the tilted distribution $\nu_{\lambda^\star}$ is simply ν with shifted natural parameters, so existing random-number generators for ν (e.g. standard library functions for Gaussian, Exponential, Gamma) can be reused directly with the tilted parameters – exponential tilting is, computationally, "free" in the sense of requiring no new sampling algorithms, only a (typically fast, per Algorithm 17.2) computation of $\lambda^\star$ and a reweighting step. This computational simplicity is a major practical advantage of exponential-family tilting over the more general (but computationally heavier) adaptive methods of Sections 19.4–19.5, whenever the underlying distribution is, in fact, in an exponential family with a tractable CGF.

Key formulas: Section 19.2

- Esscher tilt: $\nu_\lambda(dx) = e^{\lambda x - \Lambda(\lambda)}\nu(dx)$; LDP-optimal tilt $\lambda^\star$ solves $\Lambda'(\lambda^\star) = x$.
- Likelihood ratio for sums: $L = e^{-n(\lambda^\star \bar{X}_n - \Lambda(\lambda^\star))} \approx e^{-nI(x)} \approx p_n$ on the typical (tilted) event.
- Exponential family closure: Gaussian $\to$ shift mean; Exponential/Gamma $\to$ shift rate $\theta \to \theta - \lambda$.

Notation and further reading. The "Esscher transform" terminology (Section 19.2) originates in actuarial mathematics (Esscher, 1932); the same construction is called "exponential tilting," "exponential change of measure," or simply "tilting" in probability and statistics, and the "Gibbs measure" or "canonical ensemble" construction in statistical mechanics – all denote the same family $\nu_\lambda(dx) \propto e^{\lambda x}\nu(dx)$. Bucklew's *Introduction to Rare Event Simulation* provides an accessible treatment focused specifically on the exponential-tilting approach to importance sampling for sums of i.i.d. random variables.

Cross-volume connections. Exponential tilting's role in "making the rare event typical" parallels a construction in Volume III: robust optimisation under a relative-entropy ambiguity set (Section 17.3's connection) can be solved via a dual problem involving exactly the same exponential-tilting (Esscher) transform, with the dual variable $\lambda^\star$ playing the role of a Lagrange multiplier for the ambiguity-budget constraint – the "optimal tilt for importance sampling" and "optimal worst-case distribution for robust optimisation under KL ambiguity" are, formally, the same optimisation problem viewed from different angles (estimation vs. decision-making), both resolved by the identical exponential-family change of measure.

Example — Extended discussion: multivariate exponential tilting with correlation

For $\mathbf{X} \sim \mathcal{N}(\square, \square)$, the tilt $\nu_\square \propto e^{\square\cdot\mathbf{x}}\nu$ is $\mathcal{N}(\square + \square\square, \square)$ – shifts the mean by $\square\square$, *leaves* $\square$ *unchanged*. The optimal tilt for target $\mathbf{x}^\star$ solves $\square^\star = \square^{-1}(\mathbf{x}^\star - \square)$.
Worked example. Continuing Section 17.2's two-asset extended discussion ($\square = (0, 0)$, $\square$ with $\rho = 0.6$, $\mathbf{x}^\star = (-0.04, -0.04)$): $\square^\star = \square^{-1}\mathbf{x}^\star \approx -62.5(1, 1)$. Since $\square\square^\star = \mathbf{x}^\star$, $\Lambda(\square^\star) = \frac{1}{2}\square^{\star\top}\square\square^\star = \frac{1}{2}\square^\star\cdot\mathbf{x}^\star = 2.5$, matching $\Lambda^\star(\mathbf{x}^\star) = 2.5$ via $\Lambda^\star = \square^\star\cdot\mathbf{x}^\star - \Lambda(\square^\star) = 5.0 - 2.5 = 2.5$. ✓
Preserving correlation. The tilted simulation $\mathbf{X} \sim \mathcal{N}((-0.04, -0.04), \square)$ retains $\rho = 0.6$ – tilted "both down 4%" scenarios exhibit the *same* co-movement structure as the original model, just recentred. This matters for downstream analysis (Section 19.2's stress-scenario application note): the tilt changes *where* scenarios are centred, not the *relationships* between variables within them.

Example — Extended discussion: tilting fails for regularly-varying tails

Exponential tilting requires $\mathbb{E}[e^{\lambda X}] < \infty$ near $\lambda = 0$ – exactly the condition that fails for regularly-varying tails (Section 17.2's heavy-tail extended discussion). For such X, no exponential change of measure makes a rare large value typical; the natural alternative changes the *tail index itself* – simulating from a distribution with the same body but an artificially altered polynomial tail, with importance weights growing only polynomially (not exponentially) in x.
Consequence. Variance-reduction factors achievable for heavy-tailed rare-event simulation are typically far more modest than the dramatic factors of the light-tailed examples in this section – a structural limitation, not a tuning failure, that should temper expectations when this section's machinery meets heavy-tailed loss models.

19.3 Splitting Methods and Subset Simulation

Overview

Splitting methods (also called subset simulation, multilevel splitting, or RESTART) avoid the need for an explicit change of measure entirely. The rare event A is decomposed as a nested sequence of progressively less rare events $A = A_m \subset A_{m-1} \subset \cdots \subset A_0 = \Omega$, and $p = \Pr(A) = \prod_{k=1}^{m} \Pr(A_k \mid A_{k-1})$ is estimated as a product of conditional probabilities, each $O(1)$ and hence estimable with few samples – "splitting" the hard problem into a chain of easy ones.

Definition: Nested levels and conditional probabilities

Let $\Phi : \mathcal{X} \to \mathbb{R}$ be a "score function" with $A = \{\Phi(X) \geq a_m\}$ for the target threshold a_m. Choose intermediate thresholds $-\infty = a_0 < a_1 < \cdots < a_m$ defining $A_k = \{\Phi(X) \geq a_k\}$, so $A_0 = \Omega \supseteq A_1 \supseteq \cdots \supseteq A_m = A$. Then

$$p = \Pr(A_m) = \prod_{k=1}^{m} \Pr(A_k \mid A_{k-1}) = \prod_{k=1}^{m} p_k, \qquad p_k := \Pr(A_k \mid A_{k-1}).$$

If thresholds are chosen so each $p_k \approx p_0$ for some moderate p_0 (e.g. $p_0 = 0.1$), and $m = \log p / \log p_0$ levels are used, each p_k can be estimated by ordinary Monte Carlo with N samples drawn from the conditional distribution $\nu(\cdot \mid A_{k-1})$.

Example — Theorem: Variance of the multilevel estimator

Let $\hat{p}_k$ be an unbiased estimator of p_k from N (conditionally independent, given perfect conditional sampling) samples, with $\hat{p} = \prod_k \hat{p}_k$. By independence across levels (under the idealisation of exact conditional sampling),

$$\frac{\operatorname{Var}(\hat{p})}{p^2} \approx \sum_{k=1}^{m} \frac{\operatorname{Var}(\hat{p}_k)}{p_k^2} \approx \sum_{k=1}^{m} \frac{1 - p_k}{N p_k} \approx m \cdot \frac{1 - p_0}{N p_0}$$

(using $\operatorname{Var}(\hat{p}_k) = p_k(1-p_k)/N$ for a binomial proportion estimator and $p_k \approx p_0$ for all k). The relative error is $\mathrm{RE} \approx \sqrt{m(1-p_0)/(Np_0)}$ – crucially, this grows only *linearly* in $\sqrt{m} \sim \sqrt{|\log p|}$, not as $1/\sqrt{p}$ as for crude MC. Total simulation cost $\sim mN$ grows only *logarithmically* in $1/p$, an exponential improvement over crude MC's $\Omega(1/p)$ cost. *Conditional sampling in practice.* Exact sampling from $\nu(\cdot \mid A_{k-1})$ is itself generally infeasible; the standard solution (subset simulation, Au & Beck 2001) uses Markov chain Monte Carlo (MCMC), specifically a Metropolis–Hastings sampler targeting $\nu(\cdot \mid A_{k-1})$, seeded from the surviving samples of level $k-1$ (those with $\Phi(X) \geq a_{k-1}$). The level thresholds a_k themselves are often chosen *adaptively* as empirical quantiles of $\{\Phi(X_i)\}$ among the current population, ensuring $p_k \approx p_0$ automatically (Section 19.5 develops this adaptive idea further).

Example — Worked example: subset simulation for a structural reliability problem

A structure fails if a load-dependent stress $\Phi(\mathbf{X}) = \max_i \sigma_i(\mathbf{X})$ (the maximum stress over d components, each a function of random load/material variables $\mathbf{X} \in \mathbb{R}^d$) exceeds a critical threshold $a_m = \sigma_{\text{crit}}$, with target failure probability $p = \Pr(\Phi(\mathbf{X}) \geq \sigma_{\text{crit}}) \sim 10^{-6}$. Direct MC with $N = 10^4$ samples would expect 0.01 failures –

essentially never observing one. Subset simulation with $p_0 = 0.1$ needs $m = \log(10^{-6})/\log(0.1) = 6$ levels. With $N = 1000$ samples per level: level 1 uses 1000 i.i.d. samples from ν, takes a_1 as the 90th percentile of $\{\Phi(X_i)\}$ (so $\hat{p}_1 \approx 0.1$ by construction); level 2 seeds 1000 MCMC chains from the 100 samples with $\Phi(X_i) \geq a_1$, generating new samples from $\nu(\cdot \mid \Phi \geq a_1)$, and sets a_2 as their 90th percentile; and so on through level 6, where $a_6 = \sigma_{\text{crit}}$ is fixed and $\hat{p}_6$ is the empirical fraction (of the level-6 population, conditioned on level 5) exceeding σ_{crit}. Total cost: $6 \times 1000 = 6000$ samples (plus MCMC burn-in overhead) to estimate a 10^{-6} probability – versus $\gtrsim 10^8$ for crude MC to get a comparable relative error, a $\sim 10^4\times$ computational saving, scaling further (logarithmically) for even smaller target probabilities.

Relation to large deviation theory. Splitting methods are, in a sense, "model-free" importance sampling: they do not require knowledge of $\Lambda^\star$ or V, only a reasonable score function Φ and the ability to do local MCMC moves. However, large deviation theory still informs *design choices*: the natural score function Φ for a path-space rare event is often suggested by the structure of the dominant large-deviation path $\phi^\star$ (Section 18.1) – e.g. $\Phi(\text{path}) =$ "progress along $\phi^\star$" – and the number of levels m needed is, to leading order, $m \approx I(A)/(-\log p_0)$ where $I(A)$ is the LDP rate (so m scales linearly with the rate-function value, i.e. with the "depth" of the rare event in large-deviation units).

Example — Exercises

1. For $p = 10^{-9}$ and $p_0 = 0.2$, compute $m = \log p/\log p_0$. *Solution:* $\log(10^{-9}) \approx -20.72$; $\log(0.2) \approx -1.609$. $m \approx -20.72/-1.609 \approx 12.9$, so $m = 13$ levels (rounding up).
2. For $N = 2000$ samples per level, $m = 13$, $p_0 = 0.2$, compute RE $\approx \sqrt{m(1-p_0)/(Np_0)}$. *Solution:* RE $\approx \sqrt{13 \times 0.8/(2000 \times 0.2)} = \sqrt{10.4/400} = \sqrt{0.026} \approx 0.161$ – about 16% relative error with a total of 26000 samples, versus $\sim 10^{11}$ for crude MC at $p = 10^{-9}$.
3. If p_0 is increased from 0.2 to 0.5, how does m change for the same target $p = 10^{-9}$, and what is the new RE (for $N = 2000$)? *Solution:* $\log(0.5) \approx -0.693$, so $m \approx 20.72/0.693 \approx 29.9 \to 30$ levels. RE $\approx \sqrt{30 \times 0.5/(2000 \times 0.5)} = \sqrt{15/1000} \approx 0.122$ – slightly better RE, at the cost of more total samples (60000 vs. 26000) – a trade-off between number of levels and per-level difficulty.

Further perspectives. Subset simulation (Au & Beck, 2001) was developed within the structural reliability engineering community, largely independently of the probabilistic large-deviation and splitting literature (which traces to Kahn & Harris, 1951, in the context of nuclear shielding calculations – "particle splitting" for radiation transport, one of the very earliest variance-reduction techniques in Monte Carlo history). The convergence of these independently-developed traditions onto essentially the same algorithmic idea – decompose a rare event into a chain of conditionally-likely events – mirrors the convergence noted for Gärtner–Ellis (Section 17.4) and Eyring–Kramers (Section 18.4): recurring rediscovery of the same mathematical structure from different applied starting points is itself evidence of the structure's fundamental character. A practical caveat for the MCMC step within subset simulation: the Markov chain used to sample from $\nu(\cdot \mid A_{k-1})$ must be *ergodic with respect to this conditional distribution*, which can be subtle to verify or achieve when A_{k-1} has a complex (e.g. disconnected, or with narrow "necks") geometry – poor mixing within a level can introduce correlations between the "samples" at that level that violate the independence assumption underlying the variance formula, leading to underestimated uncertainty. Diagnostics for within-level mixing (e.g. effective sample size calculations) are therefore an essential, if often under-emphasised, component of a careful subset simulation implementa-

tion.

Example — Application note: subset simulation for dam overtopping risk

A dam's reservoir level is modelled as a stochastic process driven by rainfall and inflow, with "overtopping" defined as the level exceeding the dam crest, a target probability on the order of $p \sim 10^{-5}$ per year (a "1-in-100,000-year" design standard, typical for major dam safety regulations). Direct simulation of n-year reservoir-level trajectories to estimate $p \sim 10^{-5}$ would require $N \sim 10^7$-10^8 simulated years for adequate precision – computationally demanding if each simulated year involves a detailed hydrological model. Subset simulation (Section 19.3) with $\Phi =$ "maximum reservoir level reached during the simulated period" and $p_0 = 0.1$ requires $m = \log(10^{-5})/\log(0.1) = 5$ levels, each requiring $N \sim 1000$ detailed hydrological simulations (conditioned, via MCMC perturbation of rainfall/inflow scenarios, on the previous level's threshold) – a total of ~ 5000 detailed simulations, versus 10^7-10^8 for direct simulation, a $\sim 10^4\times$ computational saving directly enabling dam safety assessments that would otherwise be computationally infeasible at the required design probability. A subtlety specific to this application: the "MCMC perturbation" of a rainfall/inflow scenario (Step in the subset simulation algorithm) must respect the physical structure of the rainfall process (e.g. spatial and temporal correlation of precipitation) – naive perturbations (e.g. independently resampling each time step) can produce physically implausible "perturbed" scenarios that, while formally valid for the algorithm's mathematics, do not represent realistic conditional rainfall patterns, potentially biasing the resulting probability estimate. Designing physically-faithful MCMC moves for conditional scenario generation is, in practice, often the dominant engineering challenge in applying subset simulation to physically-based risk models, considerably more involved than the mathematics of the splitting algorithm itself.

Algorithm: subset simulation with adaptive levels

Input: target threshold a_m, score function Φ, level probability p_0 (e.g. 0.1), samples per level N.

1. Simulate N i.i.d. samples $X_1, \dots, X_N \sim \nu$; compute $\Phi(X_i)$ for each.
2. Set $a_1 =$ the $(1-p_0)$-quantile of $\{\Phi(X_i)\}$; let $S_1 = \{X_i : \Phi(X_i) \geq a_1\}$ (the "seeds," $\approx p_0 N$ of them).
3. For $k = 2, 3, \dots$: for each seed in S_{k-1}, run an MCMC chain (e.g. Metropolis–Hastings with a symmetric proposal, accepting moves only if the proposed state also satisfies $\Phi \geq a_{k-1}$) for several steps to generate a roughly-independent sample from $\nu(\cdot \mid \Phi \geq a_{k-1})$; collect N such samples total (multiple per seed, if $|S_{k-1}| < N$). Set $a_k = \min(a_m, (1-p_0)$-quantile of these N samples' Φ values$)$; let $S_k = \{$samples with $\Phi \geq a_k\}$.
4. Stop when $a_k = a_m$ (i.e. the quantile would exceed a_m, so a_m itself is used). Let $\hat{p}_k = |S_k|/N$ for the final level, and $\hat{p}_j = p_0$ for all earlier levels (by construction).
5. $\hat{p} = \prod_k \hat{p}_k$.

The "multiple samples per seed" provision in step 3 (needed because $|S_{k-1}| \approx p_0 N < N$, so generating N new samples requires, on average, $1/p_0$ samples per seed) introduces within-level correlation (samples descending from the same seed are correlated) – this is the source of the "MCMC mixing" diagnostic concern raised in Section 19.3's remarks, and is typically addressed by running each seed's chain for enough steps that consecutive samples from the same seed are approximately independent (an "thinning" or "sub-sampling" the chain), at

the cost of additional computation per seed.

Key formulas: Section 19.3

- Multilevel decomposition: $p = \prod_{k=1}^{m} p_k$, $p_k = \Pr(A_k \mid A_{k-1})$.
- RE $\approx \sqrt{m(1-p_0)/(Np_0)}$, $m \approx \log p / \log p_0$.
- Cost $O(mN)$ – logarithmic in $1/p$, vs. crude MC's $O(1/p)$.
- Levels chosen adaptively as empirical quantiles of Φ within the current population.

Notation and further reading. "Subset simulation" (Au & Beck) and "multilevel splitting" (the probability/physics tradition) denote essentially the same algorithm developed independently; "RESTART" (Repetitive Simulation Trials After Reaching Thresholds) is a closely related variant from the telecommunications-networks literature. The unifying mathematical treatment is sometimes called "sequential importance sampling with resampling" when framed within the sequential Monte Carlo literature. Rubino & Tuffin's edited volume *Rare Event Simulation using Monte Carlo Methods* surveys the splitting family of methods alongside importance sampling, with applications across telecommunications, reliability, and finance.

Cross-volume connections. Subset simulation's nested-event decomposition mirrors the hierarchical decomposition of cascading failure events in Volume IV: "the network fails completely" is decomposed into "10% of nodes fail" $\to$ "30% fail" $\to$ ... $\to$ "100% fail," each conditional step estimated via simulation – precisely Section 19.3's algorithm, applied to network failure progression rather than abstract score functions. The structural-reliability origins of subset simulation (Section 19.3's remarks) also connect to Volume II: subset simulation was developed specifically to handle exactly the kind of small-failure-probability structural reliability problems (Section 19.3's application note on dam overtopping) that Volume II's reliability theory addresses analytically for simpler (e.g. series/parallel) system structures.

Example — Extended discussion: optimal p_0 via cost-variance trade-off

Fix target RE $= \epsilon_0$. From $\text{RE}^2 = \frac{m(1-p_0)}{Np_0}$, total cost $C = mN = \frac{m^2(1-p_0)}{\epsilon_0^2 p_0} = \frac{L^2}{\epsilon_0^2} \cdot \frac{1-p_0}{\ell_0^2 p_0}$ where $L = \log(1/p)$, $\ell_0 = \log(1/p_0)$, $m = L/\ell_0$.
Numerical check. For $p = 10^{-6}$ ($L \approx 13.8$), $\epsilon_0 = 0.10$ ($L^2/\epsilon_0^2 = 19044$): at $p_0 = 0.1$ ($\ell_0 \approx 2.303$), $g = \frac{0.9}{0.1 \times 2.303^2} \approx 1.698$, $C \approx 32{,}340$. At $p_0 = 0.2$ ($\ell_0 \approx 1.609$), $g \approx 1.545$, $C \approx 29{,}425$. At $p_0 = 0.3$ ($\ell_0 \approx 1.204$), $g \approx 1.609$, $C \approx 30{,}650$. So $p_0 \approx 0.2$ minimises cost here – consistent with refinements of Au & Beck's original $p_0 \approx 0.1$ suggestion. The $\sim 10\%$ cost differences between $p_0 = 0.1$ and 0.2 are modest but non-negligible for expensive per-sample simulations.

Example — Extended discussion: inter-level correlation in subset simulation

Each level's MCMC step starts from "seed" samples that are themselves correlated – they are the surviving elites from the previous level's already-correlated chains. This inter-level correlation propagates: the conditional estimates $\hat{p}_k \approx \Pr(A_k|A_{k-1})$ are not independent, and $\text{CV}(\hat{p})$ for $\hat{p} = \prod_k \hat{p}_k$ includes between-level covariance terms not captured by this section's within-level formula alone.
Mitigation. Running multiple independent subset-simulation chains from scratch and averaging the resulting $\hat{p}$ removes inter-run correlation (though not intra-run inter-level correlation), giving a more honest variance estimate via the between-run

sample variance – a standard safeguard whenever this section's algorithm produces a final reported probability rather than an exploratory estimate.

19.4 Cross-Entropy Method for Rare-Event Estimation

Overview

The cross-entropy (CE) method is an adaptive importance sampling scheme: starting from a parametric family of proposal distributions $\{q_\theta\}$, it iteratively updates θ to minimise the Kullback–Leibler divergence $D(q^\star \| q_\theta)$ to the (unknown) zero-variance proposal $q^\star = \nu(\cdot \mid A)$ of Section 19.1 – replacing the analytic tilting of Section 19.2 with a data-driven, simulation-based search, applicable when $\Lambda^\star$ is intractable.

Definition: Cross-entropy minimisation

Given a parametric family $\{q_\theta\}_{\theta \in \Theta}$ (e.g. $q_\theta = \mathcal{N}(\theta, \square)$, varying the mean), the CE method seeks

$$\theta^\star = \arg\min_\theta D(q^\star \| q_\theta) = \arg\max_\theta \mathbb{E}_{q^\star}[\log q_\theta(X)],$$

i.e. the θ under which q_θ best matches the conditional law $q^\star = \nu(\cdot \mid A)$ in the maximum-likelihood / cross-entropy sense. Since $\mathbb{E}_{q^\star}[\cdot] = \mathbb{E}_\nu[\cdot \mid A] = \frac{1}{p}\mathbb{E}_\nu[\mathbf{1}_A(X)\cdot]$, this becomes

$$\theta^\star = \arg\max_\theta \mathbb{E}_\nu[\mathbf{1}_A(X) \log q_\theta(X)]$$

(the $1/p$ cancels in the argmax) – a weighted maximum-likelihood problem, estimable by simulation *even from the original* ν, provided A is observed at least occasionally.

Example — Theorem: Iterative CE algorithm

Since A may be too rare to observe under ν directly, the CE method bootstraps via a sequence of intermediate, less-rare target sets $A^{(0)} \supseteq A^{(1)} \supseteq \cdots \supseteq A^{(T)} = A$ (analogous to the levels of Section 19.3), defined adaptively:

1. Initialise θ_0 such that $q_{\theta_0} = \nu$ (no tilt).

2. At iteration t**:** draw $X_1, \ldots, X_N \sim q_{\theta_{t-1}}$, compute scores $\Phi(X_i)$, and set $A^{(t)} = \{\Phi \geq \gamma_t\}$ where γ_t is the empirical $(1-\rho)$-quantile of $\{\Phi(X_i)\}$ for a fixed $\rho \in (0,1)$ (e.g. $\rho = 0.1$) – ensuring $\Pr_{q_{\theta_{t-1}}}(A^{(t)}) \approx \rho$, an "achievable" target.

3. Update: solve the weighted MLE

$$\theta_t = \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{A^{(t)}}(X_i)\, \frac{\nu(X_i)}{q_{\theta_{t-1}}(X_i)}\, \log q_\theta(X_i),$$

which for exponential-family $\{q_\theta\}$ (e.g. Gaussian with mean θ) has a closed-form solution: θ_t is the (importance-weighted) sample mean of the $X_i \in A^{(t)}$.

4. Iterate until $\gamma_t \geq a_m$ (the true threshold); then use q_{θ_T} as the final IS proposal to estimate p directly.

Connection to exponential tilting. For a Gaussian base $\nu = \mathcal{N}(0, \square)$ and exponential-family proposals $q_\theta = \mathcal{N}(\theta, \square)$, the CE-optimal $\theta^\star$ in the limit $A^{(t)} \to A$ *coincides* with the exponential tilt $\square^\star\square$ of Section 19.2 in the special case where A is a half-space $\{\mathbf{c}\cdot\mathbf{x} \geq a\}$ – the CE method, for exponential-family proposals, automatically rediscovers the

LDP-optimal tilt without ever computing $\Lambda^\star$ analytically.

Example — Worked example: CE for a multivariate threshold exceedance

Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ and $A = \{\sum_{i=1}^5 X_i \geq 10\}$ (a sum of 5 standard normals exceeding 10, i.e. 2 standard deviations of the sum, since $\mathrm{Var}(\sum X_i) = 5$, $\mathrm{sd} = \sqrt{5} \approx 2.24$, so threshold ≈ 4.47 sd – $p \approx \Pr(\mathcal{N}(0,5) \geq 10) = \Pr(\mathcal{N}(0,1) \geq 4.47) \approx 3.9 \times 10^{-6}$). Take $q_\theta = \mathcal{N}(\theta \mathbf{1}, \mathbf{I}_5)$ (a common shift θ across coordinates by symmetry). Iteration: start $\theta_0 = 0$; draw $N = 1000$ samples, compute $\Phi(\mathbf{X}) = \sum X_i \sim \mathcal{N}(0,5)$, take $\gamma_1 =$ the 90th percentile ≈ 2.87 ($\approx 1.28 \times \sqrt{5}$). Update $\theta_1 =$ mean of $\sum X_i/5$ over the 100 samples with $\Phi \geq \gamma_1$, roughly $\theta_1 \approx 0.65$. Repeat: γ_2 becomes the 90th percentile of Φ under $q_{\theta_1} = \mathcal{N}(0.65\mathbf{1}, \mathbf{I}_5)$ (mean 3.25, so $\gamma_2 \approx 3.25 + 1.28\sqrt{5} \approx 6.11$), update $\theta_2 \approx 1.3$, etc. After $T \approx$ 3-4 iterations γ_t reaches 10 and $\theta_T \approx 2$ (exactly the LDP-optimal tilt for this half-space event, by the symmetry argument: the optimal tilt direction is $\mathbf{1}/\sqrt{5}$ scaled so $\mathbb{E}_{q_{\theta^\star}}[\sum X_i] = 5\theta^\star = 10 \Rightarrow \theta^\star = 2$, matching $\Lambda'(\lambda^\star)$ for the appropriate $\lambda^\star$). The final proposal $q_{\theta_T} = \mathcal{N}(2\mathbf{1}, \mathbf{I}_5)$ is then used for the production IS run – the entire tilt direction and magnitude were discovered by $\sim 4 \times 1000 = 4000$ simulations, without ever writing down $\Lambda^\star$ for this 5-dimensional problem.

Practical considerations. The CE method's main weaknesses are: (1) for highly multimodal $q^\star$ (e.g. multiple disjoint failure regions), a single parametric family q_θ may poorly approximate $q^\star$ even at convergence – mixture proposals are then needed; (2) the choice of ρ trades off convergence speed (large ρ, slow but stable) against the risk of premature convergence to a local, non-representative mode (small ρ); (3) for very high-dimensional $\mathbf{X}$ (e.g. entire SDE paths, Section 18.1), parametric families for q_θ must be chosen carefully – often parametrising a *drift shift* (cf. the Girsanov tilting of Section 19.2) rather than the joint path distribution directly.

Example — Exercises

1. For $q_\theta = \mathcal{N}(\theta, 1)$ and a sample $\{x_1, \ldots, x_N\} \subseteq A^{(t)}$ (the elite set), derive the CE update θ_t. *Solution:* The weighted log-likelihood is $\sum_i w_i \log q_\theta(x_i) = \sum_i w_i(-\frac{1}{2}(x_i - \theta)^2 + \text{const})$; maximising over θ gives $\theta_t = \frac{\sum_i w_i x_i}{\sum_i w_i}$, the weighted sample mean of the elite set.
2. If after the first CE iteration the elite set (top 10% by Φ) has sample mean $\bar{x} = 1.5$ (with weights ≈ 1 since $\theta_0 = 0$), what is θ_1? *Solution:* $\theta_1 = \bar{x} = 1.5$ directly, since the likelihood-ratio weights are all 1 at $\theta_0 = 0$.
3. Explain why, for $q_\theta = \mathcal{N}(\theta, 1)$ and a half-space event $A = \{X \geq a\}$, the CE-optimal $\theta^\star$ at convergence should satisfy $\theta^\star \approx a$ for large a. *Solution:* The zero-variance proposal $q^\star = \mathcal{N}(0,1)|_{X \geq a}$ (truncated normal) has, for large a, mean very close to a itself (the truncated-normal mean excess over a shrinks like $1/a$). The CE method, fitting $\mathcal{N}(\theta, 1)$ to approximate $q^\star$, converges to $\theta^\star \approx a$, matching the exponential tilt $\lambda^\star = a$ for $\mathcal{N}(0,1)$ (since $\Lambda(\lambda) = \lambda^2/2$, $\Lambda'(\lambda) = \lambda$).

Further perspectives. The cross-entropy method (Rubinstein, 1997, originally for combinatorial optimisation, later adapted to rare-event simulation) exemplifies a broader paradigm sometimes called "learning the importance sampling distribution" – as opposed to deriving it analytically (Section 19.2). This learning paradigm has, in recent years, been extended using neural-network-parametrised proposal families q_θ (rather than simple exponential-family distributions), trained via stochastic gradient descent on the CE objective (or related

objectives, such as the KL divergence in the opposite direction, $D(q_\theta \| q^\star)$, which has different – generally mode-seeking rather than mass-covering – properties and is the objective used in variational inference more broadly). For high-dimensional path-space rare events (e.g. entire trajectories of a complex simulator), neural proposal families can represent state-and-time-dependent drift adjustments $u_\theta(x, t)$ far more flexibly than the constant or affine tilts of classical exponential-family approaches, effectively learning an approximation to the optimal control $u^\star$ of the stochastic-control perspective mentioned in Section 19.2's remarks. This connects rare-event simulation to the broader and very active field of "neural importance sampling," with applications ranging from particle physics (rare collision events) to climate modelling (rare extreme weather simulation) – contexts where the underlying dynamics are too complex for any closed-form $\Lambda^\star$ or V, but where simulation itself (the source of training data for q_θ) remains feasible, if expensive.

Example — Application note: cross-entropy for portfolio stress scenario discovery

A multi-asset portfolio's loss $L(\mathbf{X})$ depends on $\mathbf{X} \in \mathbb{R}^{50}$ (returns of 50 risk factors) via a complex (non-linear, e.g. involving options with convex/concave payoffs) pricing function with no closed-form $\Lambda^\star$. To find the "1-in-1000-year" loss scenario ($A = \{L(\mathbf{X}) \geq \ell_{0.999}\}$), the cross-entropy method (Section 19.4) with $q_\theta = \mathcal{N}(\square, \square)$ (50-dimensional mean-shift family, covariance $\square$ fixed at the baseline) iterates: simulate $N = 5000$ scenarios from $q_{\theta_{t-1}}$, identify the elite set (top 10% by L), update θ_t as the elite set's mean. After convergence (typically 5-10 iterations for a well-behaved L), $\square_T \in \mathbb{R}^{50}$ identifies the "centre of mass" of the 50-dimensional stress scenario most associated with A – often revealing, e.g., that the dominant stress direction involves a specific combination of 5-10 risk factors moving together (a "stress mode"), even though the original 50-dimensional space offered no obvious low-dimensional structure. This dimensionality reduction – the CE method effectively performs an *adaptive, risk-weighted* principal-component-like analysis, identifying the direction(s) in factor space most associated with extreme portfolio loss – is often as valuable as the probability estimate itself: portfolio managers can use $\square_T$ to design hedges targeting specifically the identified stress mode, a far more actionable output than a single scalar probability $\hat{p}$. The caveat from Section 19.4's discussion – that a single Gaussian q_θ may poorly represent multimodal $q^\star$ – is particularly relevant here: if the portfolio contains both "long volatility" and "short volatility" positions, there may be *two* distinct stress modes (a "vol spike" mode and a "vol crush" mode) both contributing to extreme losses, and a single-mode CE fit would identify only one (typically whichever has larger probability mass in A), potentially leaving the portfolio manager unaware of the second vulnerability – motivating mixture-proposal extensions of CE for portfolios with genuinely heterogeneous risk exposures.

Algorithm: cross-entropy method, full pseudocode

Input: base distribution ν, exponential family $\{q_\theta\}$ with $q_{\theta_0} = \nu$, elite fraction ρ (e.g. 0.1), sample size N, target threshold a_m.

1. $\theta \leftarrow \theta_0$; $\gamma \leftarrow -\infty$.
2. **While** $\gamma < a_m$:
   (a) Sample $X_1, \ldots, X_N \sim q_\theta$; compute $\Phi(X_i)$ and weights $w_i = \nu(X_i)/q_\theta(X_i)$.
   (b) $\gamma \leftarrow \min(a_m, (1-\rho)$-quantile of $\{\Phi(X_i)\})$.
   (c) Elite set $E = \{i : \Phi(X_i) \geq \gamma\}$.

(d) $\theta \leftarrow$ weighted-MLE update using $\{(X_i, w_i) : i \in E\}$ (closed-form for exponential-family q_θ, per Section 19.4's exercises).

3. Final IS run: sample N' (typically $N' \gg N$) from q_θ, compute $\hat{p} = \frac{1}{N'} \sum_i \nu(X_i)/q_\theta(X_i) \mathbf{1}_{\{\Phi(X_i) \geq a_m\}}$.

A practical safeguard not always emphasised: the weighted-MLE update in step 2(d) uses weights $w_i = \nu(X_i)/q_\theta(X_i)$ from the *current* θ (not θ_0), and as θ moves away from θ_0 over iterations, these weights can themselves become heavy-tailed (the same ESS concern as Algorithm 19.1) – a common fix is to use the elite-set *empirical* distribution directly (i.e. set $w_i = 1$ for $i \in E$, ignoring the importance weights in the update step, an approximation valid when $q_{\theta_{t-1}}$ and q_{θ_t} are not too different) – a bias-variance trade-off (the unweighted update is biased but lower-variance) that practitioners tune via the elite fraction ρ and the number of iterations.

Key formulas: Section 19.4

- CE objective: $\theta^\star = \arg\min_\theta D(q^\star \| q_\theta) = \arg\max_\theta \mathbb{E}_\nu[\mathbf{1}_A \log q_\theta(X)]$.
- Iterative scheme: $A^{(t)} = \{\Phi \geq \gamma_t\}$, $\gamma_t = (1-\rho)$-quantile of Φ under $q_{\theta_{t-1}}$; $\theta_t =$ weighted MLE on elite set.
- Gaussian family: $\theta_t =$ weighted sample mean of elite set.
- Convergence: for half-space A and Gaussian q_θ, $\theta^\star$ recovers the LDP-optimal exponential tilt.

Notation and further reading. The "elite set" terminology (Section 19.4) is standard in the cross-entropy literature but is also called the "selected sample" or "top-ρ quantile sample" in related evolutionary-optimisation contexts (the cross-entropy method shares conceptual DNA with estimation-of-distribution algorithms in optimisation). Rubinstein & Kroese's *The Cross-Entropy Method* is the standard monograph, covering both the rare-event-simulation and combinatorial-optimisation applications of the method side by side – a juxtaposition that usefully highlights the method's generality beyond the probability-estimation focus of this volume.

Cross-volume connections. The cross-entropy method's "iteratively re-weight toward the rare region" logic parallels the iterative re-estimation schemes used in Volume II for reliability-growth modelling (where a system's failure-rate model is iteratively updated as new failure data accumulate, with "more weight" implicitly given to recent, more failure-relevant data) and in Volume IV for network vulnerability analysis (iteratively identifying "the most vulnerable" sub-network via repeated stress-testing and re-weighting toward identified weak points) – a general pattern of "adaptive focusing on the currently-most-informative region" that recurs whenever a fixed computational budget must be allocated toward the most decision-relevant (often, rarest) outcomes.

Example — Extended discussion: mixture proposals for multimodal targets

A single-component q_θ can fail dramatically for multimodal $q^\star$. For $A = \{|X| > 2.5\}$ under $\nu = \mathcal{N}(0,1)$ (a two-sided event with symmetric modes in $\nu(\cdot|A)$), a single-Gaussian CE fit $q_\theta = \mathcal{N}(\theta, 1)$ converges to $\theta = 0 =$ the *original, untilted* distribution by symmetry – no improvement at all.

Mixture fix. With $q_\theta = \frac{1}{2}\mathcal{N}(-\theta, 1) + \frac{1}{2}\mathcal{N}(\theta, 1)$, the CE objective is maximised at $\theta \approx 2.5$–3 (matching the target's actual modes): $\log q_\theta(-3)$ for $\theta \approx 3$ is ≈ -1.61 versus

$\log \mathcal{N}(0,1)(-3) \approx -5.42$ – a dramatically better fit. The mixture CE update classifies elite samples by $\text{sign}(x)$, then updates each component using only its assigned samples (an EM-type hard-assignment simplification).

Diagnostic. "CE converges to θ_0 (no-op) despite A being genuinely rare" is a strong signal of multimodality requiring a mixture family – distinct from Algorithm 19.1's low-ESS diagnostic (poor-but-non-trivial tilt) in that here the search itself finds *no* improving direction.

Example — Extended discussion: choosing the elite fraction ρ

The elite fraction ρ (typically 0.01–0.1) controls a bias-variance trade-off in each CE update. Small ρ: the elite set resembles samples from the true zero-variance tilt (low bias) but is small (high update variance). Large ρ: more samples reduce update variance, but the elite set is less representative of the true tail, biasing the update toward the proposal's bulk.

Adaptive choice. Setting ρ implicitly via a target elite-set size $N\rho \approx 50$–200, rather than a fixed fraction, automatically increases ρ as N grows or decreases it as the proposal sharpens – an analogue, at the level of CE iterations, of subset simulation's adaptive level-threshold selection (Section 19.3); both convert a fixed-fraction heuristic into an adaptive, sample-size-aware rule.

19.5 Adaptive Multilevel Splitting

Overview

Adaptive multilevel splitting (AMS) combines the level-based decomposition of Section 19.3 with full adaptivity in both the choice of levels *and* the resampling mechanism, achieving provably consistent estimates with minimal tuning – and, unlike basic subset simulation, AMS has been proven to satisfy a central limit theorem with an explicit asymptotic variance, enabling rigorous confidence intervals for the rare-event probability estimate.

Definition: AMS algorithm (population-based)

Maintain a population of N "particles" (sample paths or points) $X^{(1)}, \dots, X^{(N)} \sim \nu$. At each iteration:

1. Compute scores $\Phi(X^{(i)})$ for all particles; let $a = \min_i \Phi(X^{(i)})$ be the current minimum (the "worst" particle).
2. **Selection:** discard the k particles achieving the minimum score a (typically $k = 1$).
3. **Branching:** for each discarded particle, resample a replacement by taking a *surviving* particle (uniformly, or weighted) and regenerating its trajectory *from the point where it first exceeded level a* (for path-space problems) or perturbing it via an MCMC move conditioned on $\Phi \geq a$ (for static problems).
4. Record a as a new level; repeat until all particles satisfy $\Phi(X^{(i)}) \geq a_m$ (the target threshold).

The probability estimate is $\hat{p} = \prod_j \left(1 - \frac{k_j}{N}\right)$, the product over all iterations j of the surviving fraction – each iteration's surviving fraction is an unbiased estimate of the

conditional "advance probability."

Example — Theorem: Consistency and CLT for AMS

As $N \to \infty$, $\hat{p} \to p$ almost surely, and

$$\sqrt{N}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \qquad \sigma^2 = -p^2 \log p$$

(under idealised conditions: $k = 1$ discarded per iteration, and the score function Φ has a continuous distribution so ties have probability zero). The asymptotic relative variance is $\sigma^2/p^2 = -\log p = |\log p|$, so $\mathrm{RE}(\hat{p}) \approx \sqrt{|\log p|/N}$ – remarkably, this depends on p *only logarithmically*, confirming the same qualitative scaling as Section 19.3's heuristic ($m \sim |\log p|$ levels), but now with an exact, distribution-independent asymptotic constant. This universality – the asymptotic relative variance of AMS depends on the target probability p but *not* on the dimension of $\mathbf{X}$, the shape of A, or the choice of (reasonable) score function Φ – is AMS's signature theoretical advantage over both basic splitting (Section 19.3, whose variance depends on level-spacing choices) and importance sampling (Section 19.2, whose variance depends sensitively on how well the tilt matches $q^\star$).

Choice of score function and its (lack of) impact on asymptotic variance. While $\sigma^2 = -p^2 \log p$ is independent of Φ asymptotically, the *finite-N* performance and the number of iterations required do depend on Φ: a score function whose level sets $\{\Phi \geq a\}$ closely approximate the true "distance to A" (e.g. informed by the large-deviation path $\phi^\star$ of Section 18.1, when available) converges to the asymptotic regime with smaller N. Poor score functions (e.g. level sets that cut across the dominant escape path at oblique angles) can require much larger N to reach the asymptotic CLT regime, even though the limiting variance formula is unchanged.

Example — Worked example: AMS for the double-well exit problem

Return to the double-well diffusion $dX_t = -U'(X_t)dt + \sqrt{2\epsilon}\, dW_t$, $U(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$, starting at $X_0 = -1$, with target $A = \{\tau_0 \leq T\}$ (the path reaches $x = 0$, the saddle, by time T), for small ϵ where $p = \Pr(A) \approx e^{-(1/4)/\epsilon}$ is exponentially small (Section 18.1). Take $\Phi(\text{path}) = \max_{t \leq T} X_t$ (the maximum position reached) – a natural score, since reaching $x = 0$ means $\Phi \geq 0$. Initialise $N = 200$ independent paths from $X_0 = -1$. At each AMS iteration, the path with the smallest $\max_t X_t$ is discarded and replaced by cloning a surviving path from the point where it first exceeded the current level a, then continuing the simulation with fresh noise from that point. For $\epsilon = 0.05$ (recall from Section 18.4, $\mathbb{E}[\tau^\epsilon] \sim \pi\sqrt{2}\, e^5 \approx 660$, so direct simulation to $T \sim 1000$ would only occasionally see a crossing): AMS requires roughly $|\log p|/|\log(1 - 1/N)| \approx (1/(4 \times 0.05))/(1/200) \approx 5 \times 200 = 1000$ iterations to drive the level from $a \approx -1$ (start) up to $a = 0$ (target), at a total simulation cost of $O(N \times \text{iterations}) = O(2 \times 10^5)$ short path segments – versus the $O(10^6)$–$O(10^7)$ direct simulations that Section 18.4's calculation suggested would be needed for the much smaller $\epsilon = 0.02$ case, illustrating how AMS's cost scales roughly *linearly* in $|\log p|$ (hence linearly in $1/\epsilon$ for these exponential probabilities) rather than as $1/p$ itself.

Synthesis of Part 19. Across importance sampling (19.1–19.2), splitting (19.3), cross-entropy (19.4), and AMS (19.5), the unifying theme is that *rare-event probabilities of the form $p \approx e^{-c/\epsilon}$ or e^{-nI} can be estimated at computational cost polynomial in $|\log p|$ (i.e. in $1/\epsilon$ or n), rather than*

$1/p$ – turning an exponentially-hard direct-simulation problem into a tractable one, *provided* the practitioner can identify either a good tilt direction (19.2, 19.4) or a good score function (19.3, 19.5). Large deviation theory (Parts 17–18) is precisely the source of such tilts and score functions when the underlying model admits an explicit rate function or quasi-potential; when it does not, the adaptive methods of this Part (CE, AMS) discover the necessary structure empirically. Part 20 turns to a complementary question: given *data* on rare events (whether observed directly or via the simulation methods of this Part), how should one perform Bayesian statistical inference about the underlying tail parameters?

Part 20

Bayesian and Extreme Inference for Rare Events

Inferential frameworks for quantifying uncertainty about rare-event probabilities from limited data.

Example — Exercises

1. For $p = 10^{-4}$, compute $\sigma^2 = -p^2 \log p$ and $\mathrm{RE} = \sqrt{|\log p|/N}$ for $N = 500$. *Solution:* $|\log p| = 4 \log 10 \approx 9.21$. $\mathrm{RE} = \sqrt{9.21/500} \approx 0.136$ – about 13.6% relative error with $N = 500$ particles, regardless of dimension.
2. How many particles N are needed for $\mathrm{RE} = 0.05$ at $p = 10^{-4}$? *Solution:* $0.05 = \sqrt{9.21/N} \Rightarrow N = 9.21/0.0025 \approx 3684$ – roughly 4000 particles, independent of dimension or the shape of A, in stark contrast to crude MC's $N \sim 1/(p\,\mathrm{RE}^2) \approx 4 \times 10^6$.
3. For $p = 10^{-8}$, how does the required N for $\mathrm{RE} = 0.05$ change, and what does this say about AMS's scalability? *Solution:* $|\log p| = 8 \log 10 \approx 18.4$, so $N \approx 18.4/0.0025 \approx 7363$ – only roughly double the N for $p = 10^{-4}$, despite p being $10^4\times$ smaller. This logarithmic scaling is AMS's key advantage for extremely rare ($p < 10^{-10}$) events.

Further perspectives. Adaptive multilevel splitting, in the form presented here (often called the "Adaptive Multilevel Splitting" or AMS algorithm of Cérou & Guyader, 2007), is closely related to – and in some formulations identical to – sequential Monte Carlo (SMC) samplers used in Bayesian computation (where the "levels" are inverse-temperature or data-tempering schedules rather than score thresholds), again illustrating the cross-pollination between rare-event simulation and Bayesian computational methods that becomes explicit in Part 20. The exact CLT result $\sigma^2 = -p^2 \log p$, while elegant, relies on idealised assumptions (continuous score distribution, exactly one particle discarded per iteration, perfect "branching" from a uniformly-chosen surviving particle); real implementations introduce additional variance from finite-population effects and imperfect branching (e.g. when branching from a surviving particle's trajectory requires choosing a "branching time" along that trajectory, introducing correlations between offspring). Quantifying and correcting for these finite-N effects is an active area of research, with techniques borrowed from the broader particle-filter / SMC literature (e.g. effective-sample-size-based resampling triggers, to avoid excessive correlation among particles). Despite these practical subtleties, the universal $-p^2 \log p$ scaling remains the right "back of envelope" benchmark against which any rare-event simulation scheme – whether AMS, importance sampling, or subset simulation – should be compared, and a scheme that cannot approach this benchmark (after reasonable tuning) likely has a remediable algorithmic inefficiency rather than a fundamental barrier.

Example — Application note: AMS for nuclear plant safety analysis

Probabilistic safety assessments for nuclear power plants target "core damage frequency" estimates on the order of $p \sim 10^{-5}$ to 10^{-7} per reactor-year – among the

most extreme target probabilities in any engineering risk discipline, and a domain where AMS-type methods (or their close relatives, originally developed for exactly this purpose in the particle-transport / "splitting" tradition noted in Section 19.3's remarks) have a decades-long track record. For $p = 10^{-6}$, the AMS relative-error formula RE $\approx \sqrt{|\log p|/N}$ with $|\log p| = 6 \log 10 \approx 13.8$ gives, for $N = 2000$ particles, RE $\approx \sqrt{13.8/2000} \approx 0.083$ – an 8.3% relative error on a 10^{-6} probability from just 2000 "particles" (each particle representing a full plant-system simulation trajectory, itself potentially computationally expensive, so $N = 2000$ is far from "cheap" in absolute terms, but vastly cheaper than the $\sim 10^8$ direct simulations crude MC would require). The choice of score function Φ here is itself a major engineering decision: a natural choice might be a weighted combination of safety-system degradation indicators (e.g. number of failed redundant cooling pumps, reactor pressure deviation from nominal, ...), designed by domain experts to approximate "distance to core damage" – and, per Section 19.5's discussion, while the *asymptotic* variance formula $\sigma^2 = -p^2 \log p$ is independent of this choice, the *practical* convergence to that asymptotic regime (i.e. how large N must be before the formula is a good approximation) depends strongly on how well Φ's level sets align with the true "approach to core damage" geometry – making score-function design, informed by decades of plant-specific engineering judgement, as important to a successful safety assessment as the AMS algorithm itself.

Algorithm: AMS with explicit branching

Input: N particles, score Φ, target a_m.

1. Initialise N independent trajectories from ν; for path-space problems, simulate each to a fixed horizon T (or until absorption/exit).
2. Compute $\Phi^{(i)} = \Phi(\text{trajectory } i)$ for each i; let $a = \min_i \Phi^{(i)}$, $k = |\{i : \Phi^{(i)} = a\}|$ (number achieving the minimum – typically $k = 1$ for continuous Φ).
3. If $a \geq a_m$: stop; all particles have reached the target.
4. Discard the k particles with $\Phi^{(i)} = a$. For each, select (uniformly at random) a surviving particle j, identify the time τ_j at which trajectory j first reached level a (its "branch point"), copy trajectory j's history up to τ_j, and simulate a *fresh, independent* continuation from τ_j to T using new random numbers – this is the "branching" step, producing a new trajectory that agrees with j's up to level a but is independent thereafter.
5. Record $\hat{p}_{\text{step}} = (N - k)/N$ (the surviving fraction); accumulate $\hat{p} \leftarrow \hat{p} \times \hat{p}_{\text{step}}$.
6. Repeat from step 2.

Step 4's "branch point" identification – finding $\tau_j = \inf\{t : \Phi(\text{trajectory } j \text{ up to } t) \geq a\}$ – requires the score Φ to be a function of the *trajectory up to time* t, not just the final state, for path-space problems (e.g. $\Phi(\text{path}) = \max_{t \leq T} X_t$, the running maximum, as in Section 19.5's worked example, has a well-defined "first time the running max reaches level a"). For score functions that depend only on the endpoint X_T (not the path), the AMS algorithm degenerates to a form of the static (non-path) splitting of Section 19.3, with "branching" simply meaning "resample from the conditional endpoint distribution given $\Phi \geq a$" – the path-space and static formulations of AMS/splitting are thus unified by whether Φ depends on the full trajectory or only its endpoint, with the path-space case requiring the additional bookkeeping of branch-point identification.

Key formulas: Section 19.5

- AMS: discard worst particle(s), branch from a surviving particle at the current level a.
- $\hat{p} = \prod_j (1 - k_j/N)$.
- CLT: $\sqrt{N}(\hat{p} - p) \to \mathcal{N}(0, \sigma^2)$, $\sigma^2 = -p^2 \log p$.
- RE $\approx \sqrt{|\log p|/N}$ – dimension- and shape-independent, logarithmic in $1/p$.

Notation and further reading. "Adaptive multilevel splitting" (AMS) is due to Cérou & Guyader (2007); closely related algorithms appear under the names "Sequential Monte Carlo samplers" (Del Moral, Doucet & Jasra, 2006) in the Bayesian computation literature and "Diffusion Monte Carlo" / "Quantum Monte Carlo branching" in computational physics – again, the recurring theme of independent rediscovery across fields. Cérou, Guyader, Lelievre, and collaborators have produced an extensive rigorous literature on AMS's CLT and variance properties; for an accessible overview connecting AMS to the broader rare-event simulation landscape, see the relevant chapters of Rubino & Tuffin (cited under Section 19.3).

Cross-volume connections. AMS's universal $\sigma^2 = -p^2 \log p$ scaling provides a benchmark against which Volume II's reliability-block-diagram-based analytical methods (which give exact, closed-form rare-event probabilities for series/parallel/k-out-of-n systems, but do not extend to systems with complex dependence structures) can be compared when extended, via simulation, to more realistic dependent-component systems: AMS applied to a complex (non-block-diagram) reliability system should, for comparable target probabilities, achieve the *same* relative error as the analytical formulas would predict for an "equivalent" block-diagram system – a useful sanity check when validating simulation-based reliability assessments against simpler analytical benchmarks from Volume II.

Example — Extended discussion: resampling schemes and variance

AMS's branching step selects a surviving particle to clone. *Multinomial* resampling selects independently and uniformly with replacement – introducing extra variance from chance repeated/missed selections even among "equally good" survivors. *Systematic* resampling uses a single random offset plus equally-spaced increments, ensuring near-exactly-proportional selection counts, eliminating this extra variance.

When it matters. For $k = 1$ (Algorithm 19.5's standard case), multinomial and systematic resampling *coincide* – both simply select one survivor uniformly; the schemes differ only for $k > 1$, where systematic resampling *coordinates* multiple selections. This is why Section 19.5's CLT result $\sigma^2 = -p^2 \log p$ (derived for $k = 1$) is scheme-independent.

Implementation choice. Some AMS variants discard $k = 0.1N$ per iteration for efficiency (fewer, larger iterations); for $k > 1$, systematic resampling's variance reduction becomes material, partially recovering the efficiency loss multinomial resampling would otherwise incur – a genuine implementation trade-off paralleling Section 19.3's p_0-choice trade-off (more, smaller steps vs. fewer, larger ones).

Example — Extended discussion: bias and convergence rate of AMS estimators

The AMS estimator $\hat{p} = \prod(\text{surviving fractions})$ is asymptotically unbiased as $N \to \infty$ for fixed k, but exhibits $O(1/N)$ bias at finite N from the discreteness of the branching step – ties in the reaction-coordinate ranking must be broken, introducing a small systematic effect, largest when k/N is large and the reaction coordinate has many near-

ties.

Practical check. Comparing $\hat{p}$ for $k = 1$ versus $k = \lceil 0.1N \rceil$ at matched computational cost, and confirming agreement within the standard errors of this section's CLT formula, provides an empirical bias diagnostic – systematic disagreement beyond expected sampling noise signals that N is too small relative to k for the asymptotic theory to apply reliably.

20.1 Bayesian Estimation of Extreme Quantiles

Overview

Extreme quantiles – e.g. the 99.97% value-at-risk, or the 1-in-10,000-year flood level – are, by construction, far beyond the bulk of observed data. Frequentist extreme value theory (Volume I) fits a GPD or GEV tail and reports a point estimate $\hat{x}_q$; Bayesian inference instead produces a full *posterior distribution* over x_q, propagating parameter uncertainty into the quantile estimate itself – essential when, as is typical, the data are too sparse to pin down tail parameters precisely.

Definition: Posterior predictive quantile

Let $X_1, \dots, X_n$ (or, for threshold exceedances, the excesses over a threshold u) follow a GPD(ξ, σ) with unknown shape ξ and scale σ, prior $\pi(\xi, \sigma)$, and posterior $\pi(\xi, \sigma \mid \text{data}) \propto L(\xi, \sigma \mid \text{data})\pi(\xi, \sigma)$. The *posterior predictive* distribution for a future observation X_{n+1} is

$$p(x \mid \text{data}) = \int p(x \mid \xi, \sigma)\, \pi(\xi, \sigma \mid \text{data})\, d\xi\, d\sigma,$$

and the Bayesian quantile estimate $\hat{x}_q$ solves $\int_{-\infty}^{\hat{x}_q} p(x \mid \text{data}) dx = q$ – i.e. $\hat{x}_q$ is the q-quantile of the posterior predictive, automatically averaging over parameter uncertainty (a form of "model averaging" over the continuum of (ξ, σ) values).

Example — Theorem: Posterior predictive variance exceeds plug-in variance

Let $\hat{x}_q(\xi, \sigma)$ denote the GPD quantile function for fixed parameters. By the law of total variance applied to X_{n+1} conditional on $(\xi, \sigma) \sim \pi(\cdot \mid \text{data})$,

$$\text{Var}(X_{n+1} \mid \text{data}) = \mathbb{E}_{\xi,\sigma}[\text{Var}(X_{n+1} \mid \xi, \sigma)] + \text{Var}_{\xi,\sigma}(\mathbb{E}[X_{n+1} \mid \xi, \sigma]).$$

The second term – variance *due to parameter uncertainty* – is exactly the extra spread that a "plug-in" estimate (using point estimates $\hat{\xi}, \hat{\sigma}$ and ignoring their uncertainty) misses. For extreme quantiles, where $\hat{x}_q(\xi, \sigma)$ is highly sensitive to ξ (recall from Volume I that the tail index ξ controls whether moments even exist), this second term can dominate, especially for q close to 1: the posterior predictive quantile $\hat{x}_q$ can be substantially larger than the plug-in quantile $\hat{x}_q(\hat{\xi}, \hat{\sigma})$, reflecting genuine epistemic uncertainty about how heavy the tail really is.

Example — Worked example: posterior predictive 99.9% quantile, GPD with informative prior

Suppose $n = 50$ threshold exceedances (over $u = 0$) are observed, with sample mean excess $\bar{e} = 2.0$ and sample variance of excesses $s^2 = 8.0$. A method-of-moments-type prior centred near $\hat{\xi}_0 = 0.5((\bar{e}^2/s^2) - 1) = 0.5(0.5 - 1) = -0.25$ would suggest a light tail, but we instead posit a more conservative elicited prior $\xi \sim \mathcal{N}(0.3, 0.15^2)$ reflecting a moderately heavy tail (truncated to $\xi > -0.5$ for finite-mean excesses), with $\sigma \,|\, \xi$ a weakly informative Gamma centred at $\hat{\sigma} = 2$. Posterior simulation (via MCMC, e.g. random-walk Metropolis on $(\xi, \log\sigma)$) with $n = 50$ exceedances yields, say, a posterior mean $\mathbb{E}[\xi \,|\, \text{data}] \approx 0.32$ with posterior sd ≈ 0.10 (the data only modestly update the prior, as is typical with $n = 50$). The plug-in 99.9% exceedance quantile at $\hat{\xi} = 0.32, \hat{\sigma} = 2$ is $x_{0.999} = \frac{\hat{\sigma}}{\hat{\xi}}[(1 - 0.999)^{-\hat{\xi}} - 1]$ (GPD quantile formula) $= \frac{2}{0.32}[(0.001)^{-0.32} - 1] = 6.25 \times [10^{0.32 \times 3} - 1] = 6.25 \times [10^{0.96} - 1] \approx 6.25 \times (9.12 - 1) = 6.25 \times 8.12 \approx 50.8$. The posterior *predictive* quantile, integrating over $\xi \sim \mathcal{N}(0.32, 0.10^2)$ (and similarly over σ's posterior), is typically 20–40% higher – e.g. ≈ 68 – because the quantile function $x_{0.999}(\xi)$ is convex and increasing rapidly in ξ near $\xi \approx 0.3$ (by Jensen's inequality, $\mathbb{E}[x_{0.999}(\xi)] \geq x_{0.999}(\mathbb{E}[\xi])$ for convex $x_{0.999}(\cdot)$), so posterior uncertainty in ξ pushes the predictive quantile *above* the plug-in value – a systematic, quantifiable "uncertainty loading" that a responsible risk assessment should report alongside the point estimate.

Discussion. The convexity argument above is general: whenever the quantile-of-interest is a convex function of the tail parameters (typically true for $\xi > 0$ and q close to 1), Bayesian posterior predictive quantiles are systematically more conservative than plug-in estimates – and the gap widens as $q \to 1$ (more extreme) or as the posterior on ξ becomes more diffuse (less data, or genuinely uncertain physical regime). This "uncertainty loading" is not a Bayesian artefact to be corrected away; it reflects a real feature of decision-making under parameter uncertainty (cf. Volume III's treatment of robust decision-making under model ambiguity, of which Bayesian model averaging is one resolution). Section 20.2 addresses how to choose the prior $\pi(\xi, \sigma)$ – a choice that, as the example shows, can materially affect the resulting risk estimate, especially with limited data.

Example — Exercises

1. Compute the GPD quantile $x_{0.99}$ for $\xi = 0.2, \sigma = 1$. *Solution:* $x_{0.99} = \frac{\sigma}{\xi}[(0.01)^{-\xi} - 1] = 5 \times [10^{0.4} - 1] \approx 5 \times 1.512 = 7.56$.
2. Recompute for $\xi = 0.4$, same $\sigma = 1$, and compare. *Solution:* $x_{0.99} = 2.5 \times [10^{0.8} - 1] \approx 2.5 \times 5.31 \approx 13.27$ – nearly double the $\xi = 0.2$ value, illustrating the strong sensitivity of even moderate (99%) quantiles to ξ.
3. If posterior $\xi \sim \mathcal{N}(0.3, 0.05^2)$ (tighter than the worked example's 0.10), would the gap between posterior-predictive and plug-in 99.9% quantiles be larger or smaller? *Solution:* Smaller – the Jensen-gap "uncertainty loading" scales (to leading order) with $\text{Var}(\xi)$. A posterior sd of 0.05 (vs. 0.10) corresponds to $\frac{1}{4}$ the variance, so the gap would be roughly $\frac{1}{4}$ as large.

Further perspectives. The posterior-predictive-vs-plug-in comparison generalises beyond GPD quantiles to essentially *any* risk measure that is a non-linear (especially convex) function of underlying parameters: Value-at-Risk (a quantile, as discussed), Expected Shortfall / CVaR (an average of tail quantiles, also generally convex in tail-shape parameters for heavy tails), and even seemingly "linear" quantities like expected loss can exhibit posterior-

predictive inflation if the *existence* of the expectation itself depends on the parameter (e.g. $\mathbb{E}[X] = \infty$ for $\xi \geq 1$ in the GPD) – in such cases, even a small posterior probability mass on $\xi \geq 1$ can make the posterior predictive mean *literally infinite*, regardless of how small that probability mass is, a stark (if somewhat pathological) illustration of how parameter-uncertainty propagation can produce qualitatively, not just quantitatively, different answers than plug-in analysis. Practitioners typically address this by reporting posterior predictive quantiles (always finite for $q < 1$) rather than means when ξ's posterior has support on or near $\xi \geq 1$, or by imposing a prior that effectively bounds ξ away from 1 (a "soft" version of the moment-existence constraints that Volume I discusses in the frequentist context) – another instance of how prior choice (Section 20.2) and the choice of *which* summary to report are intertwined decisions.

Example — Application note: regulatory capital under parameter uncertainty

A regulatory framework requires capital to be set at the 99.5% posterior predictive quantile of annual loss, explicitly incorporating parameter uncertainty (as opposed to a plug-in 99.5% quantile at point-estimated parameters) – a requirement increasingly common in insurance regulation (e.g. Solvency II's "one-year value-at-risk" with explicit parameter-uncertainty loadings). For a GPD-tailed loss model with $n = 80$ historical exceedances, posterior $\mathbb{E}[\xi] \approx 0.28$, posterior sd ≈ 0.08, the plug-in 99.5% quantile might be $\hat{x}_{0.995}(\hat{\xi} = 0.28, \hat{\sigma}) \approx \$120\text{M}$, while the posterior predictive 99.5% quantile (via the Jensen-gap mechanism of Section 20.1) might be $\approx \$138\text{M}$ – a 15% "parameter uncertainty loading." If the institution instead had $n = 300$ historical exceedances (a longer/richer dataset) with the *same* posterior mean $\mathbb{E}[\xi] \approx 0.28$ but a tighter posterior sd ≈ 0.04 (roughly half, consistent with the $\sqrt{n}$ scaling of posterior precision for well-behaved likelihoods), the Jensen-gap loading would shrink to roughly $\frac{1}{4}$ its previous *absolute* contribution to the variance (since the gap scales with $\mathrm{Var}(\xi)$, Section 20.1's exercises), translating to perhaps a $\sim$ 7-8% loading instead of 15% – i.e. $\approx \$129\text{M}$ rather than $\$138\text{M}$. This quantifies, in capital terms, the value of additional data: roughly $\$9\text{M}$ ($\approx 138 - 129$) of "capital relief" from halving the posterior uncertainty in ξ, holding the point estimate fixed – a calculation that can directly inform decisions about investment in improved data collection or modelling (e.g. pooling data across business units via Section 20.3's hierarchical approach, precisely in order to reduce posterior uncertainty in shared tail parameters and thereby reduce this uncertainty-loading component of required capital).

Algorithm: posterior predictive quantile via MCMC

Input: exceedance data $\{e_1, \ldots, e_n\}$; prior $\pi(\xi, \sigma)$; target quantile level q.

1. Run MCMC (e.g. random-walk Metropolis, or Hamiltonian Monte Carlo for better mixing in the often-correlated $(\xi, \log\sigma)$ posterior) to obtain M posterior samples $\{(\xi^{(m)}, \sigma^{(m)})\}_{m=1}^{M} \sim \pi(\xi, \sigma \mid \text{data})$.
2. For each posterior sample, simulate one (or more) draws $X^{(m)} \sim \mathrm{GPD}(\xi^{(m)}, \sigma^{(m)})$ – this gives M draws from the posterior predictive distribution.
3. $\hat{x}_q$ = the q-quantile of $\{X^{(m)}\}_{m=1}^{M}$ (empirical quantile of the pooled posterior-predictive sample).
4. For a stable estimate of an extreme quantile (e.g. $q = 0.999$), ensure $M(1-q) \gg 1$ (e.g. $M = 10^5$ for $q = 0.999$ gives ~ 100 samples beyond $\hat{x}_q$, a reasonable but not generous effective sample size for the quantile itself) – if $M(1-q)$ is too small, either increase M or use a parametric-tail extrapolation of the empirical

posterior-predictive distribution beyond its observed range.

Step 4's caveat – that even $M = 10^5$ posterior-predictive draws give only ~ 100 samples in the extreme tail when $q = 0.999$ – is a "nested rare-event" problem: estimating an extreme quantile of a distribution that itself requires simulation (because it's a posterior predictive, an infinite mixture over the posterior) compounds the rare-event challenges of Part 19 with the Bayesian computation challenges of this Part. In practice, for very extreme q (e.g. $q = 0.9999$), the GPD's *analytic* quantile formula (Section 20.1's definition box) applied to each posterior sample $(\xi^{(m)}, \sigma^{(m)})$ – giving M values $x_q^{(m)} = x_q(\xi^{(m)}, \sigma^{(m)})$ directly, then taking $\hat{x}_q$ = posterior mean (or median) of $\{x_q^{(m)}\}$ – avoids the nested-simulation issue entirely, at the cost of answering a subtly different question ("the posterior mean of the q-quantile" rather than "the q-quantile of the posterior predictive distribution"); the two coincide only if $x_q(\xi, \sigma)$ happens to be linear in (ξ, σ), which it is not (Section 20.1's Jensen-gap discussion) – so the choice between these two operationally-different "Bayesian quantile" definitions should be made deliberately, not by computational convenience alone.

Key formulas: Section 20.1

- Posterior predictive: $p(x|\text{data}) = \int p(x|\xi,\sigma)\pi(\xi,\sigma|\text{data})\, d\xi\, d\sigma$.
- GPD quantile: $x_q = \frac{\sigma}{\xi}[(1-q)^{-\xi} - 1]$.
- Law of total variance: $\text{Var}(X_{n+1}|\text{data}) = \mathbb{E}[\text{Var}(X|\xi,\sigma)] + \text{Var}(\mathbb{E}[X|\xi,\sigma])$.
- Jensen gap: for convex $x_q(\xi,\sigma)$, posterior predictive quantile $\geq$ plug-in quantile at posterior means.

Notation and further reading. "Posterior predictive distribution" is standard Bayesian terminology (see, e.g., Gelman et al.'s *Bayesian Data Analysis*); in the extreme-value-specific literature, Coles' *An Introduction to Statistical Modeling of Extreme Values* provides an accessible treatment of Bayesian GPD/GEV inference, including the posterior-predictive-quantile approach of Section 20.1, alongside the frequentist methods that form Volume I's primary focus – the two approaches share the same underlying GPD/GEV models and differ chiefly in how parameter uncertainty is propagated to quantile estimates.

Cross-volume connections. The posterior-predictive-vs-plug-in distinction directly extends Volume I's frequentist GPD/GEV quantile estimation (which produces point estimates and, via the delta method or profile likelihood, confidence intervals) to a fully Bayesian framework that propagates parameter uncertainty into the quantile itself rather than reporting it as a separate confidence interval around a point estimate – the two approaches (frequentist CI around $\hat{x}_q(\hat{\theta})$ vs. Bayesian posterior predictive $\hat{x}_q$) often give numerically similar *intervals* but conceptually different *point estimates*, a distinction risk practitioners moving between Volume I's and this Part's frameworks should keep clearly in mind.

Example — Extended discussion: prior predictive checks

A *prior predictive check* simulates $\theta \sim \pi(\theta)$, then $X \sim p(\cdot|\theta)$, and compares simulated data to plausible real-world ranges – a sanity check independent of any observed data, complementing posterior predictive checks.

Worked example. For $\xi \sim \mathcal{N}(0.3, 0.15^2)$ (truncated $\xi > -0.5$), $\sigma \sim \text{Gamma}(4, 2)$ (mean 2): conditional on $\xi = 0.6$ (a heavy-tail draw, $\sim 15\%$ of the prior's mass exceeds $\xi = 0.5$), $x_{0.99}(0.6, 2) = 3.33 \times [10^{1.2} - 1] \approx 49.4$. Simulating $M = 1000$ draws and examining the *unconditional* prior-predictive 99.9% quantile asks: "does this seem

plausible *a priori*, before any data?"
What failure reveals. If the prior predictive 99.9% quantile is implausibly large (e.g. exceeding total assets), this signals prior mis-specification *independent of* whether data would later correct it – crucial for new product lines (Section 20.2) where little data may exist for a long time. PC priors' "prefer $\xi = 0$" philosophy can be understood, via this check, as ensuring the prior predictive concentrates on "moderate" outcomes by construction – a prior predictive check empirically verifies this for any chosen prior.

Example — Extended discussion: profile likelihood as a frequentist cross-check

The Bayesian credible interval for an extreme quantile x_p can be cross-checked against a frequentist profile-likelihood interval: fix $x_p = c$, re-parametrise $(\xi, \sigma) \to (\xi, c)$ via the GPD quantile formula, maximise the profile log-likelihood $\ell_{\text{prof}}(c)$ over ξ, and invert $2[\ell_{\text{prof}}(\hat{c}) - \ell_{\text{prof}}(c)] \le \chi^2_{1,1-\alpha}$.
Interpretation. With a weakly-informative prior (e.g. Jeffreys, Section 20.2), the two intervals typically agree closely for moderate n. Substantial disagreement – especially for small n_j in a hierarchical setting (Section 20.3) – is usually attributable to the prior's regularising effect on ξ, which the profile likelihood cannot replicate, rather than to a computational error: distinguishing "the prior is doing real work" from "something is wrong" is the main diagnostic value of this comparison.

20.2 Prior Elicitation for Tail Parameters

Overview

Tail parameters such as the GPD shape ξ are notoriously difficult to estimate from data alone – the "small sample" problem is endemic to extreme value statistics, since by definition only a handful of observations inform the tail. Prior elicitation – formally encoding expert judgement or physical constraints into $\pi(\xi, \sigma)$ – is therefore not optional decoration but a central, often dominant, component of Bayesian tail inference.

Definition: Elicitation via quantile matching

A common elicitation strategy: ask an expert for their best-guess values (or credible intervals) of two or more *quantiles* x_{q_1}, x_{q_2} (e.g. "the 1-in-100-year level" and "the 1-in-10,000-year level"), which are operationally meaningful, rather than asking directly about ξ, σ, which are not. The elicited (x_{q_1}, x_{q_2}) (point values, or full distributions reflecting the expert's own uncertainty) are then mapped to an implied prior on (ξ, σ) via the GPD quantile relations

$$x_{q_i} = u + \frac{\sigma}{\xi}\left[(1-q_i)^{-\xi} - 1\right], \quad i = 1, 2,$$

solved (or, for distributional inputs, propagated via simulation) for (ξ, σ).

Example — Theorem: Sensitivity of ξ-posteriors to the prior, and Jeffreys' prior

For the GPD with n exceedances, the likelihood $L(\xi, \sigma) = \prod_{i=1}^{n} \sigma^{-1}(1 + \xi e_i/\sigma)^{-1/\xi - 1}$ (with e_i the excesses) is, for ξ near the boundary of moment existence (ξ near $0.5, 1$,

etc.) or for small n, often quite flat in ξ – multiple ξ values fit the data nearly equally well, so the posterior $\pi(\xi \mid \text{data}) \propto L(\xi,\sigma)\pi(\xi,\sigma)$ remains close to the prior $\pi(\xi)$ in shape. *Jeffreys' prior* $\pi_J(\xi,\sigma) \propto |\det \mathbf{I}(\xi,\sigma)|^{1/2}$ (the square root of the determinant of the Fisher information matrix) is a popular "objective" choice that is invariant under reparametrisation, but for the GPD, $\pi_J(\xi,\sigma) \propto \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2}$ for $\xi > -1/2$, which is itself far from flat and *informative* about ξ – there is, in a precise sense, no fully "uninformative" prior for tail-index estimation; every choice encodes assumptions, and the choice should be made consciously and transparently.
Penalised complexity (PC) priors. An alternative, increasingly standard in applied extreme value modelling, is the PC prior, which penalises departure from a base model ($\xi = 0$, the exponential/Gumbel case) via an exponential prior on a divergence-based distance from $\xi = 0$: $\pi_{\text{PC}}(\xi) \propto e^{-\theta\, d(\xi,0)}$ for a tuning parameter θ and KL-based distance d. This formalises "Occam's razor for tail shape" – absent strong evidence, prefer the simpler exponential-tail model – while still allowing the data to override the prior given sufficient evidence.

Example — Worked example: eliciting a prior for coastal flood levels

A coastal engineer elicits from domain experts: "the 1-in-100-year flood level is most likely around 3.5m above mean sea level, with a 90% credible range of $[2.8, 4.5]$m; the 1-in-10,000-year level is most likely around 6m, with a much wider 90% range of $[4, 12]$m" (reflecting greater uncertainty about the far tail, as is realistic). Treating both elicited quantiles as distributed on the log scale (to respect positivity) and propagating through the GPD quantile formula via Monte Carlo (sample $(x_{0.01}, x_{0.0001})$ from their elicited distributions, solve for (ξ,σ) for each sampled pair, and collect the resulting (ξ,σ) samples as an empirical prior), one obtains an implied prior with, say, $\mathbb{E}[\xi] \approx 0.15$ and a 90% prior interval for ξ of roughly $[-0.05, 0.40]$ – a prior that is informative (concentrated away from very heavy tails $\xi > 0.5$, reflecting the experts' implicit judgement that the far-tail uncertainty, while large in absolute terms, is not *unbounded*) but not dogmatic (still allows ξ up to 0.4). Combined with $n = 30$ years of tide-gauge exceedance data via the likelihood, the posterior $\pi(\xi,\sigma \mid \text{data})$ will reflect a genuine synthesis of expert judgement and observational evidence – and the elicitation exercise itself, by forcing experts to commit to quantile statements at *two* different return periods, has effectively elicited information about ξ (which controls how quantiles *scale* between return periods) that a single-quantile elicitation could not.

Discussion. Quantile-matching elicitation has the major advantage of asking experts about quantities they can reason about directly (return levels) rather than abstract shape parameters, but it requires care: eliciting only a *single* quantile under-determines (ξ,σ) (a one-dimensional constraint on a two-dimensional parameter), so at least two quantiles (at well-separated return periods, to maximise information about ξ) are needed, and the resulting implied prior on (ξ,σ) should always be *checked* – plotted, and its implications for other quantiles examined – rather than accepted as a black box. Section 20.3 addresses the complementary situation where data *are* available, but only in small quantities at each of several related sites or time periods – motivating hierarchical pooling.

Example — Exercises

1. Given elicited quantiles $x_{0.9} = 10, x_{0.99} = 25$ ($u = 0$), solve for (ξ,σ). *Solution:* Let $z = 10^{\xi}$. Dividing the two quantile equations gives $2.5 = \frac{z^2-1}{z-1} = z + 1 \Rightarrow z = 1.5 \Rightarrow$

$\xi = \log_{10} 1.5 \approx 0.176$. Then $\sigma = 10\xi/(z-1) \approx 10 \times 0.176/0.5 \approx 3.52$.
2. Using these (ξ, σ), compute the implied $x_{0.999}$. *Solution:* $x_{0.999} = 20 \times [10^{0.528} - 1] \approx 20 \times 2.373 \approx 47.5$ – roughly $1.9\times$ the $x_{0.99}$ value, a moderate escalation consistent with $\xi \approx 0.18$.
3. If the expert's $x_{0.99}$ were revised to 40 (keeping $x_{0.9} = 10$), what happens to ξ? *Solution:* The ratio $2.5 \to 4$, so $z + 1 = 4 \Rightarrow z = 3 \Rightarrow \xi = \log_{10} 3 \approx 0.477$ – a much heavier implied tail, showing how sensitively ξ depends on *ratios* between far-apart quantiles.

Further perspectives. Prior elicitation for tail parameters connects to a broader methodological literature on "expert elicitation" in risk analysis (e.g. the Sheffield Elicitation Framework, structured protocols for combining multiple experts' judgements, addressing both individual biases – overconfidence is endemic in quantile elicitation, with experts systematically underestimating the width of their own credible intervals – and aggregation across experts with potentially correlated information sources). A specific bias relevant to tail elicitation: experts asked for "1-in-N-year" event magnitudes often anchor on recently-experienced events, leading to elicited distributions that underweight the possibility of events larger than anything in living memory – precisely the events for which Bayesian tail modelling is most needed, and precisely the events that quantile-matching elicitation (Section 20.2) is least able to correct for, since it relies on the same experts' judgement. Calibration training (showing experts their own track record on past elicitation exercises, ideally across many different quantities) and structured "consider the opposite" prompts (explicitly asking "what would have to be true for the actual value to be $2\times$ your estimate?") are standard de-biasing techniques, but the fundamental challenge – that genuinely unprecedented tail events are, by definition, outside any expert's direct experience – means that elicited priors for extreme tails should always be treated as one (important, but not sole) input, ideally combined with physically-based constraints (e.g. from climate models, in the flood example) that do not rely solely on historical pattern-matching.

Example — Application note: prior elicitation for a new product line

An insurer launching a new product line (e.g. cyber-risk insurance, with essentially no internal historical loss data) must set initial reserves using almost entirely elicited priors. Following the quantile-matching approach of Section 20.2, actuaries elicit from cyber-security experts: "a 1-in-20-year cyber loss event for a typical insured client is around \$2M" and "a 1-in-200-year event is around \$15M" (a $7.5\times$ ratio over a $10\times$ change in return period). Using the quantile-matching formula (Section 20.2's exercises, with $z = 10^{\xi}$ and the ratio of quantiles $= z^{(\log 200/\log 20)}$... for return periods differing by a factor k in probability, i.e. $q_2 = 1 - 1/200, q_1 = 1 - 1/20$, the ratio $(1-q_2)/(1-q_1) = 20/200 = 0.1$, so $z = 10^{\xi}$ relates to the quantile ratio via $\frac{x_{q_2}}{x_{q_1}} = \frac{z^{(\log(1/(1-q_2))/\log(1/(1-q_1)))}-1}{z^0 \ldots}$, a more involved relation than the equal-ratio case of the worked example since the return-period ratio here is 10, not the implicit ratio of the worked example) – regardless of the precise algebra, the qualitative point is that a $7.5\times$ jump in loss for a $10\times$ jump in return period implies a *relatively light* tail (ξ modestly positive, perhaps $\xi \approx 0.1$-0.15), since a heavier tail ($\xi \to 1$) would imply a much larger ratio (recall from Section 20.2's worked example that even $\xi \approx 0.18$ over a $100\times$ return-period ratio gave a $1.9\times$ loss ratio – for a $10\times$ return-period ratio, a $7.5\times$ loss ratio would actually correspond to a *larger* ξ, since the same ξ produces a smaller loss-ratio over a smaller return-period ratio). As the new product line accumulates its own loss history (even just 5-10 years of data), the hierarchical pooling framework

of Section 20.3 – treating the new line as one "source" pooled with related lines (e.g. other technology/liability lines with longer histories) – provides a principled mechanism for the elicited prior to be gradually superseded by the line's own emerging data, without requiring an arbitrary "switch-over" point between "prior-based" and "data-based" reserving.

Algorithm: simulation-based prior construction from elicited quantiles

Input: elicited distributions for x_{q_1}, x_{q_2} (e.g. $\mathcal{N}$ or Lognormal, reflecting the expert's own uncertainty about each quantile); number of prior samples M.

1. For $m = 1, \ldots, M$: sample $(x_{q_1}^{(m)}, x_{q_2}^{(m)})$ from the elicited distributions (independently, or with elicited correlation if the expert's uncertainties about the two quantiles are believed to be correlated – often they are, e.g. if the expert's overall "the tail is heavier/lighter than typical" belief affects both).
2. Solve the two GPD quantile equations for $(\xi^{(m)}, \sigma^{(m)})$ (via the algebraic formula of Section 20.2's exercises, or numerically via 2D Newton's method if $u \neq 0$ complicates the closed form).
3. Discard samples with $\xi^{(m)} < -1/2$ (infinite-variance excursions beyond the typically-modelled range) or other domain violations.
4. The resulting $\{(\xi^{(m)}, \sigma^{(m)})\}$ is an empirical (sample-based) prior; for use in MCMC (Algorithm 20.1), either (a) fit a smooth parametric prior (e.g. bivariate Normal on $(\xi, \log\sigma)$) to these samples, or (b) use the samples directly as a "particle prior" within a sequential Monte Carlo scheme.

Option (b) – a particle/empirical prior – has the advantage of faithfully representing whatever shape the elicitation-implied prior actually has (which may be quite non-Gaussian, e.g. skewed or even multimodal if the elicited quantile distributions are themselves skewed), without the smoothing/approximation error of fitting a parametric form; its disadvantage is that standard MCMC samplers expect a prior density evaluable pointwise, which an empirical/particle prior does not directly provide (requiring either a kernel density estimate, introducing its own approximation, or a switch to particle-based posterior approximation methods entirely) – a genuine methodological choice, not merely an implementation detail, between fidelity to the elicitation and compatibility with standard computational tools.

Key formulas: Section 20.2

- Quantile-matching elicitation: solve $x_{q_i} = u + \frac{\sigma}{\xi}[(1-q_i)^{-\xi} - 1]$, $i = 1, 2$, for (ξ, σ).
- Jeffreys prior for GPD: $\pi_J(\xi, \sigma) \propto \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2}$, $\xi > -1/2$.
- PC prior: $\pi_{\text{PC}}(\xi) \propto e^{-\theta\, d(\xi,0)}$, penalising departure from $\xi = 0$.

Notation and further reading. "Prior elicitation" is a substantial sub-field of Bayesian statistics in its own right; O'Hagan et al.'s *Uncertain Judgements: Eliciting Experts' Probabilities* surveys elicitation methodology broadly (the "Sheffield Elicitation Framework" mentioned in Section 20.2's remarks is detailed there). For extreme-value-specific elicitation, Coles & Tawn's earlier papers (1996, on elicitation for extreme value models) remain foundational, predating but anticipating much of the quantile-matching approach presented in Section 20.2.

Cross-volume connections. Prior elicitation's reliance on expert judgement about extreme, rarely-observed quantiles connects to Volume IV's treatment of "unknown unknowns" in

network resilience – network topologies and failure modes not yet observed, but whose plausibility experts may be able to assess qualitatively even without quantitative historical precedent. The PC-prior philosophy ("absent strong evidence, prefer the simpler model," Section 20.2's remarks) also echoes Volume III's treatment of model parsimony under ambiguity: both frameworks formalise a preference for simpler models that can nonetheless be overridden by sufficient data, a "soft Occam's razor" appearing in both the prior-construction (this Part) and ambiguity-set-construction (Volume III) contexts.

Example — Extended discussion: comparing Jeffreys, PC, and elicited priors numerically

Jeffreys. $\pi_J(\xi) \propto (1+\xi)^{-1}(1+2\xi)^{-1/2}$: at $\xi = 0$, value 1; at $\xi = 0.3$, ≈ 0.608; at $\xi = 0.5$, ≈ 0.471 – a mild preference for smaller ξ.
PC. Using $\mathcal{I}(0) = 2$ (Appendix F), $d(\xi, 0) \approx |\xi|$, giving $\pi_{\text{PC}}(\xi) \propto e^{-\theta|\xi|}$ (Laplace). For $\theta = 5$: at $\xi = 0.3$, $e^{-1.5} \approx 0.223$; at $\xi = 0.5$, $e^{-2.5} \approx 0.082$. PC penalises $\xi = 0.5$ about $0.471/0.082 \approx 5.7\times$ more strongly than Jeffreys – a concrete number for "PC priors are more committed to simplicity."
Elicited. Section 20.2's elicited prior, $\mathbb{E}[\xi] \approx 0.15$, 90% interval $\approx [-0.05, 0.40]$ (so $\tau \approx 0.14$): density ratios at $\xi = 0.3, 0.5$ relative to $\xi = 0.15$ are $\approx 0.563, 0.044$ – centred at 0.15, not 0.
Synthesis. Three genuinely different beliefs: Jeffreys mildly agnostic, PC actively favours $\xi = 0$, elicited actively favours $\xi \approx 0.15$. A sensitivity analysis reporting posteriors under all three – Section 20.4's "report the spread" philosophy applied to priors rather than models – reveals how much the final estimate depends on this foundational choice.

Example — Extended discussion: a hierarchical prior across multiple business lines

When K business lines each have their own shape parameter ξ_k, a natural extension places $\xi_k \sim \mathcal{N}(\mu_\xi, \tau_\xi^2)$ with weakly-informative hyperpriors on μ_ξ, τ_ξ – echoing Section 20.3's hierarchical structure, now applied to the shape parameter rather than the scale.
Coupling. The "elicited prior" of this section becomes the prior on the *population mean* μ_ξ (one elicitation informs all K lines), while τ_ξ – estimated across lines – governs how strongly each $\hat{\xi}_k$ is pooled toward μ_ξ. Lines with little data are shrunk strongly toward μ_ξ; data-rich lines are shrunk little – the "prefer $\xi = 0$" PC-prior philosophy of this section thus operates at the population level, while individual-line estimates follow Section 20.3's data-driven shrinkage.

20.3 Hierarchical Models for Pooling Rare-Event Data

Overview

When tail data are available from multiple related sources – different weather stations, different bond issuers, different production lines – but each individually has too few exceedances for reliable tail estimation, hierarchical (multilevel) Bayesian models allow *partial pooling*: each source gets its own tail parameters, but these are themselves drawn from a shared population-level distribution, so information flows between sources without forcing them to be identical.

Definition: Hierarchical GPD model

For J sources (sites, assets, ...), each with n_j exceedances $\{e_{j,1}, \dots, e_{j,n_j}\}$:

$$e_{j,i} \,|\, \xi_j, \sigma_j \sim \text{GPD}(\xi_j, \sigma_j), \qquad \xi_j \,|\, \mu_\xi, \tau_\xi^2 \sim \mathcal{N}(\mu_\xi, \tau_\xi^2), \qquad \log \sigma_j \,|\, \mu_\sigma, \tau_\sigma^2 \sim \mathcal{N}(\mu_\sigma, \tau_\sigma^2),$$

with hyperpriors on the population-level parameters $(\mu_\xi, \tau_\xi^2, \mu_\sigma, \tau_\sigma^2)$. The τ parameters control the degree of pooling: $\tau \to 0$ forces $\xi_j \equiv \mu_\xi$ for all j (*complete pooling*, treating all sources as identical), while $\tau \to \infty$ lets each ξ_j be estimated independently from its own n_j exceedances (*no pooling*). The hierarchical model lets the data determine τ, achieving *partial pooling* that interpolates between these extremes.

Example — Theorem: Shrinkage and the James–Stein phenomenon for tail indices

For the simplified case of Gaussian "data" $\hat{\xi}_j \sim \mathcal{N}(\xi_j, v_j)$ (e.g. $\hat{\xi}_j$ the MLE from source j with known sampling variance v_j, an approximation valid for moderately large n_j) and $\xi_j \sim \mathcal{N}(\mu_\xi, \tau^2)$, the posterior mean is the precision-weighted average

$$\mathbb{E}[\xi_j \,|\, \text{data}] = \frac{\hat{\xi}_j / v_j + \mu_\xi / \tau^2}{1/v_j + 1/\tau^2} = \hat{\xi}_j - \underbrace{\frac{v_j}{v_j + \tau^2}}_{\text{shrinkage factor}} (\hat{\xi}_j - \mu_\xi).$$

Sources with small n_j (large v_j) are shrunk heavily toward the population mean μ_ξ; sources with large n_j (small v_j) are barely shrunk, retaining their individual signal. This is the Bayesian analogue of the James–Stein estimator, and it provably reduces aggregate mean-squared error across all J sources simultaneously compared to the no-pooling MLEs $\hat{\xi}_j$ – a striking and initially counter-intuitive result (Stein's paradox): pooling *unrelated-seeming* estimates improves each one, on average, provided $J \geq 3$.
Estimating τ empirically. τ^2 itself has a posterior informed by the *spread* of the $\hat{\xi}_j$ across sources, relative to their individual sampling variances v_j: if the $\hat{\xi}_j$ vary more than v_j alone would predict, τ^2 is estimated to be large (little pooling justified – genuine heterogeneity); if the $\hat{\xi}_j$ are more similar than v_j alone would predict, τ^2 is estimated to be small (strong pooling – sources are likely similar, and apparent differences are mostly noise).

Example — Worked example: pooling tail indices across $J = 8$ insurance lines

Eight lines of business each report GPD shape MLEs $\hat{\xi}_j$ from claims-exceedance data, with n_j ranging from 15 (sparse lines) to 200 (mature lines), giving sampling variances roughly $v_j \approx 0.6/n_j$ (a standard GPD asymptotic approximation for $\hat{\xi}$ near 0): v_j ranges from 0.04 (small $n_j = 15$) to 0.003 (large $n_j = 200$). Suppose the $\hat{\xi}_j$ values are $\{0.45, -0.05, 0.30, 0.55, 0.10, 0.25, 0.60, 0.15\}$ with $n_j = \{15, 200, 40, 20, 150, 80, 15, 100\}$ respectively. The hierarchical model estimates, say, $\hat{\mu}_\xi \approx 0.22$ (a precision-weighted grand mean, dominated by the high-n_j lines) and $\hat{\tau}^2 \approx 0.04$ (moderate heterogeneity). For line 1 ($\hat{\xi}_1 = 0.45$, $n_1 = 15$, $v_1 \approx 0.04$), the shrinkage factor is $v_1/(v_1 + \tau^2) = 0.04/0.08 = 0.5$, so $\mathbb{E}[\xi_1 \,|\, \text{data}] \approx 0.45 - 0.5(0.45 - 0.22) = 0.45 - 0.115 \approx 0.335$ – shrunk halfway toward the population mean. For line 2 ($\hat{\xi}_2 = -0.05$, $n_2 = 200$, $v_2 \approx 0.003$), the shrinkage factor is $0.003/0.043 \approx 0.07$, so $\mathbb{E}[\xi_2 \,|\, \text{data}] \approx -0.05 - 0.07(-0.05 - 0.22) \approx$

$-0.05 + 0.019 \approx -0.031$ – barely shrunk, since the mature line's own data are highly informative. The net effect: small lines' tail-risk capital requirements are informed by the broader portfolio's experience (reducing the chance that a small line with an unlucky $\hat{\xi}_j = 0.6$ from $n = 15$ observations triggers an unjustifiably extreme capital charge), while large lines retain their own well-estimated tail behaviour.

Discussion. Hierarchical pooling is especially valuable in catastrophe and operational risk modelling, where individual business units, geographic regions, or asset classes each have limited loss history but plausibly share some common tail-generating mechanism (e.g. macroeconomic shocks, climate drivers, or shared supply chains – the network structure of Volume IV). The model's hyperparameters (μ_ξ, τ_ξ) themselves become objects of interest, characterising the "population" tail behaviour from which individual units' tails are draws – directly relevant to Section 20.4's treatment of model uncertainty, since the hierarchical model is itself one of several plausible structures (full pooling, no pooling, partial pooling, or pooling with covariates) whose relative support the data should determine.

Example — Exercises

1. For source j with $\hat{\xi}_j = 0.5$, $v_j = 0.02$, and $\hat{\mu}_\xi = 0.2, \hat{\tau}^2 = 0.03$, compute the shrinkage factor and posterior mean. *Solution:* Shrinkage $= v_j/(v_j + \tau^2) = 0.02/0.05 = 0.4$. Posterior mean $= 0.5 - 0.4(0.5 - 0.2) = 0.38$.
2. For $v_j \to 0$ (essentially infinite data), what happens? *Solution:* Shrinkage $\to 0$; posterior mean $\to \hat{\xi}_j$ – no shrinkage, the source's own precise estimate is retained.
3. For $v_j \to \infty$ (essentially no data), what happens? *Solution:* Shrinkage $\to 1$; posterior mean $\to \hat{\mu}_\xi$ – complete shrinkage to the population mean, consistent with the no-pooling/complete-pooling interpolation.

Further perspectives. The James–Stein phenomenon (Stein, 1956; James & Stein, 1961) was originally a startling result about estimating the means of $J \geq 3$ independent normal populations – startling because it showed that the seemingly "obviously optimal" individual MLEs are *inadmissible* (uniformly improvable) once $J \geq 3$, a result with no analogue for $J \leq 2$. The hierarchical Bayesian framing presented in Section 20.3 makes the phenomenon far less mysterious: shrinkage simply reflects genuine Bayesian updating given a (data-estimated) population distribution, and "improvement for all J sources simultaneously" is a statement about *aggregate* (summed or averaged) risk, not about any individual source necessarily being better estimated (a source whose true ξ_j happens to be far from μ_ξ will, in fact, be *worse* estimated by the shrunk estimator than by its own MLE – the aggregate improvement is an average over the population of sources, some of which "give" and others "receive"). This nuance matters in practice: a business-line manager whose own line's $\hat{\xi}_j$ gets shrunk toward a portfolio-wide $\hat{\mu}_\xi$ may reasonably object if their line is genuinely different from the portfolio average – the hierarchical model's response to this objection is precisely the *posterior for* τ^2: if the data support genuine heterogeneity (large $\hat{\tau}^2$), shrinkage is automatically reduced, and the model's recommendation degrades gracefully toward "trust your own data" exactly when that is the data-supported conclusion.

Example — Application note: hierarchical pooling across geographic regions for flood risk

A national flood insurance program covers $J = 50$ regions (e.g. river catchments), each with its own historical flood-exceedance record, ranging from $n_j = 10$ years (newly-monitored catchments) to $n_j = 150$ years (long-monitored major rivers). A

naive "separate model per region" approach would produce extremely noisy $\hat{\xi}_j$ for the $n_j = 10$ catchments (sampling variance $v_j \propto 1/n_j$, so $15\times$ larger sampling variance than the $n_j = 150$ catchments). The hierarchical model (Section 20.3) pools across all 50 regions, with the population distribution $\mathcal{N}(\mu_\xi, \tau^2)$ potentially further structured by covariates (e.g. regional geology, climate zone) – a "hierarchical regression" extension in which μ_ξ itself depends on covariates, $\mu_{\xi,j} = \beta_0 + \beta_1 \cdot (\text{climate zone}_j) + \dots$, allowing the "population mean" toward which each region shrinks to itself vary systematically with known regional characteristics, rather than being a single national constant. For a newly-monitored catchment ($n_j = 10$) in a climate zone with many well-monitored *similar* catchments, this covariate-informed shrinkage target $\mu_{\xi,j}$ can be far more informative than a national average μ_ξ – effectively, the model "borrows strength" preferentially from *climatically similar* catchments, a refinement of the basic shrinkage mechanism (Section 20.3's exercises) that becomes increasingly valuable as the number of pooled sources J grows and meaningful covariate structure becomes estimable. This covariate-augmented hierarchical structure is, in essence, a Bayesian regression model for tail parameters – combining the partial-pooling benefits of Section 20.3 with the explanatory power of region-specific covariates, and representing the current state of practice for large-scale, geographically-distributed catastrophe risk pooling.

Algorithm: Gibbs sampling for the hierarchical GPD model

Input: per-source exceedances $\{e_{j,i}\}$; hyperpriors on $(\mu_\xi, \tau_\xi^2, \mu_\sigma, \tau_\sigma^2)$.

1. Initialise all $\xi_j, \sigma_j, \mu_\xi, \tau_\xi^2, \mu_\sigma, \tau_\sigma^2$ (e.g. from separate per-source MLE fits, and their empirical mean/variance for the hyperparameters).
2. **For each MCMC iteration:**
   (a) For each j: update (ξ_j, σ_j) via a Metropolis step targeting $\propto L_j(\xi_j, \sigma_j) \times \mathcal{N}(\xi_j; \mu_\xi, \tau_\xi^2) \times \text{LogNormal}(\sigma_j; \mu_\sigma, \tau_\sigma^2)$ (the per-source likelihood times the current population-level prior) – this is "almost" Algorithm 20.1's per-source MCMC, but with a data-informed (rather than fixed) prior.
   (b) Update $\mu_\xi \mid \{\xi_j\}, \tau_\xi^2$: if the hyperprior on μ_ξ is conjugate (e.g. Normal), this is a closed-form Gibbs draw from $\mathcal{N}(\frac{\sum_j \xi_j/\tau_\xi^2 + \mu_0/\tau_0^2}{J/\tau_\xi^2 + 1/\tau_0^2}, (J/\tau_\xi^2 + 1/\tau_0^2)^{-1})$ (precision-weighted combination of the data $\{\xi_j\}$ and the hyperprior $\mathcal{N}(\mu_0, \tau_0^2)$).
   (c) Update $\tau_\xi^2 \mid \{\xi_j\}, \mu_\xi$: if conjugate (e.g. Inverse-Gamma hyperprior), closed-form Gibbs draw from the corresponding Inverse-Gamma posterior.
   (d) Similarly for $\mu_\sigma, \tau_\sigma^2$.

The alternation between "update each source's parameters given the current population distribution" (step (a)) and "update the population distribution given all sources' current parameters" (steps (b)-(d)) is the defining structure of hierarchical-model Gibbs sampling, and it is this alternation that produces the shrinkage behaviour of Section 20.3's theorem: at each iteration, sources with extreme $\hat{\xi}_j$ relative to the current (μ_ξ, τ_ξ^2) are pulled toward μ_ξ (via the prior term in step (a)), while (μ_ξ, τ_ξ^2) themselves adapt to the overall pattern of $\{\xi_j\}$ (steps (b)-(c)) – the MCMC chain converges to a joint posterior in which both levels have adapted to each other, exactly the "partial pooling" equilibrium described analytically (for the Gaussian-approximation case) in Section 20.3's theorem.

Key formulas: Section 20.3

- Hierarchical model: $e_{j,i}|\xi_j, \sigma_j \sim \text{GPD}(\xi_j, \sigma_j)$, $\xi_j|\mu_\xi, \tau_\xi^2 \sim \mathcal{N}(\mu_\xi, \tau_\xi^2)$.
- Shrinkage: $\mathbb{E}[\xi_j|\text{data}] = \hat{\xi}_j - \frac{v_j}{v_j+\tau^2}(\hat{\xi}_j - \mu_\xi)$.
- $\tau \to 0$: complete pooling ($\xi_j \equiv \mu_\xi$); $\tau \to \infty$: no pooling ($\xi_j = \hat{\xi}_j$).
- James–Stein: partial pooling reduces aggregate MSE across $J \geq 3$ sources.

Notation and further reading. "Hierarchical Bayesian models" (also "multilevel models" or "random effects models" in the frequentist-adjacent mixed-models literature) are covered comprehensively in Gelman & Hill's *Data Analysis Using Regression and Multilevel/Hierarchical Models*. The "James–Stein" terminology (Section 20.3) refers to James & Stein's 1961 paper; Efron & Morris's subsequent expository papers (e.g. "Stein's Paradox in Statistics," Scientific American, 1977) remain among the most accessible introductions to the shrinkage phenomenon for non-specialist audiences.

Cross-volume connections. Hierarchical pooling across "sources" (regions, business lines, asset classes) is the statistical-inference counterpart of Volume IV's network-aggregation perspective: just as a network's system-level behaviour emerges from – but is not simply the sum of – its individual components' behaviours (with shared structure inducing correlations), a portfolio's aggregate tail-risk parameters emerge from – but are not simply the average of – its individual sources' parameters, with the hierarchical model's (μ_ξ, τ_ξ) playing a role analogous to a network's "common factor" or "systemic" parameters in Volume IV's cascading-failure models, a connection made fully explicit in Section 20.5's common-factor case study.

Example — Extended discussion: the non-centred parametrisation in detail

Centred parametrisation problem. $\xi_j \sim \mathcal{N}(\mu_\xi, \tau_\xi^2)$ directly. When τ_ξ is small (strong pooling), ξ_j and μ_ξ are highly correlated in the posterior – both must move together to maintain $\xi_j \approx \mu_\xi$. Gibbs sampling (Algorithm 20.3) moves slowly along this narrow, correlated ridge.

Non-centred fix. Write $\xi_j = \mu_\xi + \tau_\xi \tilde{\xi}_j$, $\tilde{\xi}_j \sim \mathcal{N}(0,1)$ *independent of* μ_ξ, τ_ξ *a priori*. The prior on $\tilde{\xi}_j$ no longer depends on the hyperparameters; coupling to (μ_ξ, τ_ξ) now flows only through the likelihood. When data are weakly informative (small n_j – exactly where pooling matters most), this coupling is weak, and $\tilde{\xi}_j, \mu_\xi, \tau_\xi$ become nearly independent in the posterior – the opposite of the centred case.

Quantitative check. For $n_j = 15$, $\tau_\xi^2 \approx 0.04$ (Section 20.3's worked example), prior precision $1/\tau_\xi^2 = 25$ is comparable to the likelihood precision for $n_j = 15$ – the centred $\xi_j|\cdot$ conditional is prior-dominated, inducing strong correlation with (μ_ξ, τ_ξ). The non-centred parametrisation (per Papaspiliopoulos, Roberts & Sköld's foundational analysis) typically reduces this correlation substantially in precisely this small-n_j, small-τ_ξ regime – the regime most common throughout Part 20's small-sample tail-estimation settings.

Example — Extended discussion: an explicit shrinkage-factor formula

For the normal-normal hierarchical model $\xi_j|\mu_\xi, \tau_\xi \sim \mathcal{N}(\mu_\xi, \tau_\xi^2)$ with likelihood precision $1/s_j^2$ for $\hat{\xi}_j$, the posterior mean is the classical shrinkage formula $B_j\hat{\xi}_j + (1-B_j)\mu_\xi$, $B_j = \frac{1/s_j^2}{1/s_j^2 + 1/\tau_\xi^2}$ – $B_j \to 1$ as $n_j \to \infty$, $B_j \to 0$ as $n_j \to 0$.

Numeric check. With $n_j = 15$, $\tau_\xi^2 \approx 0.04$ (this section's worked example) and an illustrative $s_j^2 \approx 0.0045$, $B_j = \frac{1/0.0045}{1/0.0045+1/0.04} = \frac{222}{222+25} \approx 0.90$ – substantial shrinkage even at $n_j = 15$, quantifying the qualitative "small n_j, strong pooling" statement that motivated the non-centred parametrisation earlier in this section.

20.4 Model Uncertainty and Bayesian Model Averaging in Tail Inference

Overview

Even after fixing a prior (Section 20.2) and a pooling structure (Section 20.3), a fundamental question remains: *which tail model*? GPD vs. lognormal vs. Weibull tails can fit the bulk of available data almost indistinguishably while implying wildly different extreme quantiles. Bayesian model averaging (BMA) addresses this by treating the model choice itself as an unknown to be integrated over, rather than selected once and for all.

Definition: Bayesian model averaging

Given a set of candidate models $\{M_1, \dots, M_K\}$ (e.g. different tail distributions, or different threshold choices u) with priors $\Pr(M_k)$ and marginal likelihoods $p(\text{data} \mid M_k) = \int L(\text{data} \mid \theta_k, M_k)\pi(\theta_k \mid M_k)\, d\theta_k$, the posterior model probabilities are $\Pr(M_k \mid \text{data}) \propto p(\text{data} \mid M_k)\Pr(M_k)$, and the BMA posterior predictive for a quantity of interest Δ (e.g. a return level) is

$$p(\Delta \mid \text{data}) = \sum_{k=1}^{K} p(\Delta \mid \text{data}, M_k)\ \Pr(M_k \mid \text{data}).$$

Example — Theorem: BMA variance decomposition and the cost of model uncertainty

By the law of total variance over the discrete mixture,

$$\operatorname{Var}(\Delta \mid \text{data}) = \underbrace{\sum_k \Pr(M_k \mid \text{data})\operatorname{Var}(\Delta \mid \text{data}, M_k)}_{\text{within-model (parameter) uncertainty}} + \underbrace{\sum_k \Pr(M_k \mid \text{data})\big(\mathbb{E}[\Delta \mid \text{data}, M_k] - \mathbb{E}[\Delta \mid \text{data}]\big)^2}_{\text{between-model uncertainty}}.$$

The second term is the genuinely *new* contribution of model averaging beyond Section 20.1's within-model parameter uncertainty: it quantifies the spread in point estimates *across* models, weighted by their posterior plausibility. When candidate models agree closely on $\mathbb{E}[\Delta \mid M_k]$ (e.g. all give similar median estimates but differ in tail shape), this term may be small even if the models' implied *extreme* quantiles diverge sharply – a reminder that the variance decomposition should be examined separately for each quantity Δ of interest, not just at one summary level.

Sensitivity to $\Pr(M_k)$. Unlike within-model Bayesian inference, where the prior $\pi(\theta \mid M_k)$ is "washed out" by enough data, the model-level prior $\Pr(M_k)$ does *not* vanish in the same way when models are non-nested and the data cannot, even asymptotically, discriminate perfectly between them (e.g. GPD with $\xi = 0.05$ vs. a lognormal tail can be nearly indistinguishable on any finite, even large, sample of moderate exceedances, while implying different extrapolated 99.99% quantiles). BMA results for far-tail quantities should therefore always be reported *together with* the sensitivity of

$\Pr(M_k \mid \text{data})$ to the choice of $\Pr(M_k)$.

Example — Worked example: GPD vs. lognormal tail for operational loss data

Operational risk loss data ($n = 120$ losses above a reporting threshold) are fit under two models: M_1 =GPD(ξ, σ) and M_2 =lognormal(μ, τ^2) restricted to the threshold-exceedance region. Both fit the bulk of the 120 losses with similar log-likelihoods (say, $\log L_1 = -410.2$, $\log L_2 = -410.8$ at posterior means – nearly indistinguishable), and with weakly informative priors the (Laplace-approximated) marginal likelihoods give posterior model probabilities $\Pr(M_1 \mid \text{data}) \approx 0.55$, $\Pr(M_2 \mid \text{data}) \approx 0.45$ – genuinely ambiguous. The two models' 99.97% quantile estimates, however, diverge: M_1 (with posterior $\hat{\xi} \approx 0.35$, a Pareto-type power-law tail) gives $\hat{x}^{(1)}_{0.9997} \approx \$85$M, while M_2 (lognormal, lighter than power-law in the extreme tail) gives $\hat{x}^{(2)}_{0.9997} \approx \$48$M – a factor of ~ 1.8 difference driven entirely by tail *shape* assumptions that the data cannot strongly discriminate. The BMA estimate is $0.55 \times 85 + 0.45 \times 48 \approx \$68.4$M – not simply an "average" in any naive sense, but the posterior-probability-weighted mixture, whose *spread* (the between-model variance term above) directly quantifies the dollar magnitude of model-structure uncertainty: a risk manager reporting "\$68M" without the BMA spread (here, roughly $\pm$\$20M from the between-model term alone) would dramatically understate the true uncertainty in the capital requirement.

Discussion. BMA does not "solve" model uncertainty – it makes it explicit and quantifiable, converting an unacknowledged modelling choice (which often defaults, in practice, to whichever model an analyst happens to favour) into a transparent, data-informed weighting with an associated uncertainty band. For regulatory and high-stakes risk applications, reporting the between-model variance term is often more decision-relevant than the point BMA estimate itself: it directly answers "how much does our risk estimate depend on a modelling choice we cannot fully justify from data alone?" – precisely the question that robust decision-making frameworks (Volume III) are designed to address when such ambiguity cannot be fully resolved by more data.

Example — Exercises

1. If $\Pr(M_1|\text{data}) = 0.7, \Pr(M_2|\text{data}) = 0.3$, $\mathbb{E}[\Delta|M_1] = 100, \mathbb{E}[\Delta|M_2] = 150$, $\text{Var}(\Delta|M_1) = 400, \text{Var}(\Delta|M_2) = 900$, compute the BMA mean and between-model variance. *Solution:* BMA mean $= 0.7 \times 100 + 0.3 \times 150 = 115$. Between-model term $= 0.7(100 - 115)^2 + 0.3(150 - 115)^2 = 157.5 + 367.5 = 525$.
2. Compute the within-model term and total variance. *Solution:* Within-model $= 0.7 \times 400 + 0.3 \times 900 = 550$. Total $= 550 + 525 = 1075$; nearly half ($525/1075 \approx 49\%$) comes from model uncertainty alone.
3. If $\Pr(M_1) = \Pr(M_2) = 0.5$ instead, recompute the BMA mean. *Solution:* $0.5 \times 100 + 0.5 \times 150 = 125$ – a $\sim 9\%$ change purely from a different model-weighting judgement, with no change in data or model fits.

Further perspectives. Bayesian model averaging's main practical bottleneck is computing marginal likelihoods $p(\text{data} \mid M_k)$, which involve integrating the likelihood over the *entire* parameter space weighted by the prior – a high-dimensional integral that is notoriously sensitive to the prior's tails (a prior with even slightly heavier tails than the likelihood's effective support can dominate the marginal likelihood integral in ways that have little to do with model fit, the "Lindley/Jeffreys paradox"). Modern approaches often sidestep marginal-likelihood computation entirely via predictive criteria: e.g. stacking (Yao et al., 2018), which

chooses model weights to optimise leave-one-out predictive performance directly, sidestepping the prior-sensitivity issues of marginal likelihoods while still producing a weighted combination with the same BMA-style variance-decomposition interpretation. For the tail-risk applications of this Part, predictive stacking is often preferable precisely because the marginal-likelihood approach can give models with poorly-specified (even if "proper") priors on tail parameters disproportionate weight or disproportionate penalty, in ways unrelated to their actual out-of-sample tail-quantile performance – which is, after all, the quantity of genuine interest. The variance-decomposition framework of this section remains valid regardless of which weighting scheme (marginal-likelihood-based BMA, stacking, or other) produces the model weights $\Pr(M_k \mid \text{data})$ (or their predictive analogues).

Example — Application note: model uncertainty disclosure in a risk report

A risk report for senior management and regulators presents the 99.9% aggregate loss quantile not as a single number but as a small table: the BMA point estimate ($\$68$M, per Section 20.4's worked example), the within-model 90% credible interval ($\$68$M $\pm\$15$M, say, reflecting parameter uncertainty within each model), and the between-model spread (the two component models' point estimates, $\$85$M and $\$48$M, with their posterior weights $0.55/0.45$). This three-part disclosure directly operationalises the variance decomposition of Section 20.4: a reader can see at a glance that (a) the headline number is a model-weighted average, not a single model's output; (b) even *within* the favoured model, there is substantial parameter uncertainty; and (c) the *choice* between models, which the data only weakly discriminate ($55\%/45\%$), drives a $\sim \$37$M $(85 - 48)$ swing in the point estimate – larger than either model's within-model credible interval width. Senior management, presented with this decomposition, might reasonably ask: "what would it take to discriminate between M_1 and M_2 more decisively?" – a question that the marginal-likelihood / Bayes-factor framework can, in principle, answer via a *value-of-information* calculation (how much would the posterior model probabilities shift, and how much would the resulting BMA quantile change, under additional hypothetical data of various sizes) – connecting Section 20.4's static model-averaging framework to a dynamic, decision-theoretic question about *which uncertainties are worth resolving*, a question that recurs whenever model uncertainty is large enough to materially affect a decision (here, capital adequacy) but small enough that targeted data collection might meaningfully reduce it.

Algorithm: Bayesian model averaging via posterior simulation

Input: candidate models $M_1, \dots, M_K$, each with its own MCMC posterior samples $\{(\theta_k^{(m)})\}_{m=1}^{M}$ (from K separate runs of Algorithm 20.1-type MCMC); posterior model probabilities $\Pr(M_k \mid \text{data})$ (estimated via, e.g., bridge sampling, thermodynamic integration, or other marginal-likelihood estimators – a substantial computational topic in its own right, not detailed here).

1. For $m = 1, \dots, M_{\text{total}}$ (a large total budget): draw $k \sim \text{Categorical}(\Pr(M_1|\text{data}), \dots, \Pr(M_K|\text{data}))$, then draw $\theta^{(m)} \sim \{\theta_k^{(j)}\}$ (uniformly from model k's posterior samples), then draw $\Delta^{(m)} \sim p(\Delta \mid \theta^{(m)}, M_k)$ (the predictive distribution under the sampled model and parameters).
2. The pooled $\{\Delta^{(m)}\}$ is a sample from the BMA posterior predictive $p(\Delta \mid \text{data}) = \sum_k p(\Delta|\text{data}, M_k) \Pr(M_k|\text{data})$.
3. Compute any desired summary (mean, quantiles, variance decomposition per Section 20.4's theorem) directly from $\{\Delta^{(m)}\}$, optionally stratified by which k

was drawn for each m (enabling direct estimation of the within- vs. between-model variance components).

This "mixture sampling" algorithm is conceptually simple once posterior samples for each model are available, but its accuracy is entirely contingent on the accuracy of $\Pr(M_k \mid \text{data})$ – if these model weights are mis-estimated (e.g. due to the marginal-likelihood sensitivities discussed in Section 20.4's remarks), the resulting BMA posterior predictive inherits that error directly (step 1's Categorical draw uses these weights verbatim). A useful robustness check: re-run the algorithm with $\Pr(M_k|\text{data})$ perturbed (e.g. $\pm 50\%$ relative changes, renormalised) and report how much the resulting summaries (especially extreme quantiles of Δ) change – a direct, simulation-based sensitivity analysis complementing the analytic between-model variance decomposition.

Key formulas: Section 20.4

- BMA: $p(\Delta|\text{data}) = \sum_k p(\Delta|\text{data}, M_k)\Pr(M_k|\text{data})$, $\Pr(M_k|\text{data}) \propto p(\text{data}|M_k)\Pr(M_k)$.
- Variance decomposition: $\text{Var}(\Delta|\text{data}) = \sum_k \Pr(M_k|\text{data})\,\text{Var}(\Delta|\text{data}, M_k) + \sum_k \Pr(M_k|\text{data})(\mathbb{E}[\Delta|M_k] - \mathbb{E}[\Delta])^2$.
- Within-model + between-model uncertainty; the latter does not vanish with more data if models are non-nested and indistinguishable in the limit.

Notation and further reading. "Bayesian model averaging" is surveyed comprehensively in Hoeting et al.'s 1999 *Statistical Science* review "Bayesian Model Averaging: A Tutorial"; for the "stacking" alternative mentioned in Section 20.4's remarks, see Yao, Vehtari, Simpson & Gelman, "Using Stacking to Average Bayesian Predictive Distributions" (Bayesian Analysis, 2018). The "Lindley/Jeffreys paradox" (marginal-likelihood sensitivity to prior tails) is discussed in most graduate Bayesian texts under "Bayes factors" and their pathologies.

Cross-volume connections. Bayesian model averaging's between-model variance term is the probabilistic-inference analogue of Volume III's robust optimisation under model ambiguity: both frameworks confront the situation "multiple models are consistent with available data, and they disagree about the quantity of interest" – BMA resolves this via probability-weighted averaging (requiring posterior model probabilities), while robust optimisation resolves it via worst-case analysis over the model set (requiring no probabilities over models, only a specification of which models are "plausible"). The choice between these two resolutions – average over models (BMA) vs. optimise against the worst model (robust) – is itself a decision-theoretic choice depending on whether the decision-maker's loss function rewards being "right on average" or "never catastrophically wrong," a framing Volume III develops at length.

Example — Extended discussion: information criteria as alternatives to marginal likelihoods

WAIC: $-2(\sum_i \log \bar{p}(y_i) - \sum_i \text{Var}_\theta(\log p(y_i|\theta)))$, computable directly from MCMC samples – the second term penalises model complexity via posterior sensitivity of predictions.

Schematic illustration. For Section 20.4's GPD-vs-lognormal comparison, suppose $\text{WAIC}_1 = 824.1$ (GPD), $\text{WAIC}_2 = 825.3$ (lognormal), difference 1.2 ± 3.5 (SE) – well within one SE, so WAIC *agrees* with the marginal-likelihood-based conclusion ($0.55/0.45$, "not decisively distinguishable"). Agreement between frameworks in-

creases confidence the conclusion isn't a prior-sensitivity artefact (marginal likelihoods are more prior-sensitive than WAIC/LOO, per the Lindley/Jeffreys concern).
Pseudo-BMA. Substituting WAIC/LOO-based weights $w_k \propto e^{-\text{WAIC}_k/2}$ for $\Pr(M_k|\text{data})$ in Algorithm 20.4 retains the variance-decomposition interpretation, but the between-model component now reflects *predictive accuracy* differences rather than marginal-likelihood differences (which conflate accuracy with prior-volume effects) – arguably better targeting what risk managers actually care about, at the cost of computing the full per-observation log-likelihood matrix.

Example — Extended discussion: the automatic Occam's razor of marginal likelihoods

The marginal likelihood $p(y|M) = \int p(y|\theta, M)\pi(\theta|M)\, d\theta$ automatically penalises extra parameters: each additional free parameter spreads the prior over a larger volume, diluting the average likelihood unless the data strongly favour the extra flexibility. The Savage–Dickey ratio formalises this for nested models: the Bayes factor favouring the simpler model equals the ratio of posterior to prior density of the extra parameter at the simple model's fixed value – if the posterior concentrates there no more than the prior already did, the data give no support to the extra complexity.
Connection to WAIC/LOO. WAIC's complexity penalty is a *local*, sample-based approximation to this global, prior-volume effect – both express "favour simplicity unless justified", but the marginal likelihood's penalty depends on the entire prior while WAIC's depends only on posterior-predictive sensitivity at the observed data, explaining occasional disagreement under unusually diffuse or informative priors.

20.5 Case Study: Catastrophe Risk Modelling

Overview

This closing case study integrates the chapter's themes – Bayesian quantile estimation (20.1), prior elicitation (20.2), hierarchical pooling across regions (20.3), and model averaging over tail forms (20.4) – into a single worked catastrophe risk model for hurricane losses, illustrating how the pieces combine in a realistic (if simplified) end-to-end analysis, and tying back to the large-deviation and rare-event-simulation themes of Parts 17–19.

Definition: The integrated model structure

A simplified hurricane loss model for $J = 4$ coastal regions combines: (i) a hierarchical GPD model (Section 20.3) for per-event regional losses $L_{j,i} \,|\, \xi_j, \sigma_j \sim \text{GPD}(\xi_j, \sigma_j)$ with $\xi_j \sim \mathcal{N}(\mu_\xi, \tau^2)$; (ii) elicited priors (Section 20.2) for (μ_ξ, τ) from meteorological experts, reflecting physical constraints on storm intensity distributions; (iii) two candidate dependence structures between regions – independent regional losses (M_{indep}) vs. a common "storm severity" factor inducing correlated regional losses (M_{common}) – compared via BMA (Section 20.4); and (iv) the resulting posterior predictive distribution for *aggregate annual loss* $\Lambda_{\text{agg}} = \sum_{j=1}^{4} L_j$, whose extreme quantiles (the "1-in-250-year aggregate loss," a standard regulatory capital benchmark) are the ultimate quantity of interest.

Example — Theorem: Why dependence structure dominates aggregate tail risk

For aggregate loss $\Lambda_{\text{agg}} = \sum_j L_j$ with each L_j heavy-tailed (GPD, $\xi_j > 0$), the tail behaviour of Λ_{agg} depends critically on the dependence structure: under M_{indep}, the sum of J independent heavy-tailed variables has a tail dominated by the *single largest* L_j (the "one big jump" principle: $\Pr(\Lambda_{\text{agg}} > x) \sim \sum_j \Pr(L_j > x)$ as $x \to \infty$, for subexponential distributions), so the aggregate tail index is $\xi_{\text{agg}} \approx \max_j \xi_j$. Under M_{common}, a shared severity factor Z (e.g. "this year's hurricane season is unusually intense") induces *positive dependence*: all L_j tend to be large simultaneously, and the aggregate tail can be *heavier* than $\max_j \xi_j$ would suggest under independence – the "common factor" effectively adds an extra layer of large-deviation cost shared across all regions, analogous to the network-resilience effects of Volume IV where correlated component failures produce system-level tail risk far exceeding what independent-component analysis predicts.

This is, again, a contraction-principle phenomenon (Section 17.5): $\Lambda_{\text{agg}} = f(L_1, \dots, L_J)$ for $f = \text{sum}$, and the rate function for Λ_{agg} is obtained by minimising the *joint* rate function of $(L_1, \dots, L_J)$ subject to $\sum_j L_j = \lambda$ – under independence, this minimisation is "cheap" only via one large L_j (others near their means); under positive dependence (common factor), the joint rate function itself is lower along the "all large together" direction, making that route to $\sum_j L_j = \lambda$ relatively more probable.

Example — Worked example: 1-in-250-year aggregate loss under both dependence models

Suppose the hierarchical fit (Section 20.3) gives posterior mean $\xi_j \approx \{0.25, 0.30, 0.20, 0.35\}$ for the four regions (after partial pooling toward $\hat{\mu}_\xi \approx 0.27$), with regional scale parameters implying individual 99.6% (1-in-250-year) regional loss quantiles of roughly $\{\$2.0\text{B}, \$1.5\text{B}, \$1.0\text{B}, \$0.8\text{B}\}$. Under M_{indep}, simulating $\Lambda_{\text{agg}} = \sum_j L_j$ with independent regional losses gives a 1-in-250-year aggregate quantile of, say, $\hat{\Lambda}^{\text{indep}}_{0.996} \approx \$3.2\text{B}$ – driven mostly by whichever single region happens to have an extreme year (the "one big jump"), with the other three contributing close to their typical (sub-extreme) values. Under M_{common}, with a moderate common factor (e.g. a shared lognormal multiplier Z with $\text{sd}(\log Z) = 0.3$ applied to all regional severities in a given year), the same 1-in-250-year aggregate quantile rises to $\hat{\Lambda}^{\text{common}}_{0.996} \approx \$4.8\text{B}$ – 50% higher, because the "extreme year" scenario now involves *all four* regions simultaneously experiencing elevated losses, not just one. Suppose the BMA posterior model probabilities (from historical data on simultaneous multi-region loss years) are $\Pr(M_{\text{common}} \mid \text{data}) \approx 0.65$, $\Pr(M_{\text{indep}} \mid \text{data}) \approx 0.35$ (the data show some, but not overwhelming, evidence of co-movement). The BMA aggregate quantile estimate is then *not* simply $0.65 \times 4.8 + 0.35 \times 3.2 \approx \$4.24\text{B}$ – because BMA mixes *distributions*, not quantiles directly, and the 0.996-quantile of a mixture is generally *not* the probability-weighted average of the component quantiles (mixture quantiles depend on the full shape of each component near that quantile) – but a full posterior predictive simulation, drawing the dependence structure itself from $\{M_{\text{indep}}, M_{\text{common}}\}$ with the stated probabilities at each simulated year, yields the correct BMA quantile, which in this case comes out close to $\$4.0\text{B}$ – substantially closer to the common-factor estimate, reflecting how the heavier tail of M_{common} dominates the upper quantiles of the mixture even at 35% posterior weight on M_{indep} (a manifestation of the "one heavy tail dominates a mixture's extreme quantiles" principle, the BMA analogue of the one-big-jump principle above).

Synthesis of Volume V. This case study closes the volume's arc: Part 17 established the rate-function language ($I = \Lambda^\star$, $D(\rho\|\nu)$) for quantifying "how rare is rare"; Part 18 extended this to dynamic, path-dependent rare events and metastability (quasi-potentials, Eyring–Kramers); Part 19 showed how to *compute* such probabilities efficiently when they are too small for direct simulation (tilting, splitting, AMS); and Part 20 addressed the genuinely statistical question of *inferring* the underlying tail parameters – and the uncertainty in that inference – from finite data, culminating in a multi-region catastrophe model where prior elicitation, hierarchical pooling, dependence-structure model averaging, and posterior predictive quantile estimation all interact. The recurring lesson across all four parts is that rare-event probabilities are exponentially sensitive to model assumptions (tail index ξ, dependence structure, quasi-potential barrier height) in ways that bulk-data fit alone cannot resolve – a theme that connects directly to Volume III's robust optimisation under model ambiguity, Volume IV's network-level cascading failure analysis (where the "common factor" effect above reappears as contagion), and Volume VI's treatment of antifragility, where exposure to – rather than mere estimation of – rare events becomes the central design consideration.

Example — Exercises

1. If three regions had $\xi_j = 0.25$ and the fourth had $\xi_4 = 0.6$, what does "one big jump" predict for ξ_{agg} under M_{indep}? *Solution:* $\xi_{\text{agg}} \approx \max_j \xi_j = 0.6$ – the aggregate tail is dominated entirely by region 4's heavy tail.
2. Under M_{common}, would ξ_{agg} be larger or smaller than 0.6? *Solution:* Potentially larger – the common factor amplifies all regions simultaneously, including region 4's already-heavy tail; the interaction can produce an aggregate tail heavier than either component alone (tail dependence compounding).
3. Why could a regulator using only M_{indep} be significantly under-capitalised? *Solution:* If the true dependence structure includes a common factor (as $\Pr(M_{\text{common}}|\text{data}) \approx 0.65$ suggests is more likely than not), the true 1-in-250-year loss is closer to \$4–4.8B rather than M_{indep}'s \$3.2B – a 25–50% shortfall precisely in the highest-stakes scenario.

Further perspectives. The "common factor" dependence structure used in this case study is the simplest member of a broad family of models for dependent extremes, generally organised around the concept of *tail dependence*: the limiting conditional probability that one variable is extreme given that another is extreme. The bivariate (and multivariate) extreme value copulas studied in Volume I provide a more flexible – and more rigorously asymptotically-justified – framework than the simple common-factor model used here, allowing, for instance, *asymmetric* tail dependence (region A's extremes predict region B's extremes more strongly than vice versa, plausible if A's storms often "feed into" conditions affecting B) and dependence structures that vary by quantile (weak dependence in the bulk, strong dependence only in the extreme tail, a pattern commonly observed empirically and termed "tail dependence without bulk dependence"). The integrated catastrophe model of this section should be understood as an illustrative simplification of a methodology that, in full generative practice, combines: physically-based catastrophe simulation engines (generating thousands of synthetic storm "years" from meteorological models, providing the dependence structure directly from physics rather than a parametric copula); the hierarchical Bayesian tail-parameter estimation of Section 20.3 applied to each region's loss-given-storm distribution; and the BMA-over-dependence-structures framework of Section 20.4 applied to compare the physically-simulated dependence against simpler parametric alternatives – the full synthesis representing the current state of the art in catastrophe risk modelling, of which this volume has traced the essential mathematical DNA from first principles (rate functions, Section 17) through to final application (Section 20.5).

Example — Application note: reinsurance structuring under the integrated model

The integrated catastrophe model of Section 20.5 directly informs reinsurance purchasing decisions: a reinsurance treaty attaching at \$3B and exhausting at \$5B (i.e. covering losses in the layer $[\$3\text{B}, \$5\text{B}]$) would, under M_{indep} alone, appear to cover essentially the entire "tail" beyond the 1-in-250-year point ($\hat{\Lambda}^{\text{indep}}_{0.996} \approx \$3.2\text{B}$, just above the attachment point, with the \$5B exhaustion point representing a much rarer event under independence). Under M_{common}, however, the 1-in-250-year point itself ($\approx \$4.8\text{B}$) is *within* the layer, and the BMA-blended estimate ($\approx \$4.0\text{B}$) sits roughly in the middle of the layer – meaning the treaty, sized based on M_{indep} alone, would leave the insurer exposed to a materially larger share of the 1-in-250-year loss than intended. A reinsurance structure informed by the full BMA posterior predictive (rather than either point estimate alone) might instead attach at \$2.5B and exhaust at \$6B – wider, to account for the genuine uncertainty in *where* the 1-in-250-year point lies – at a correspondingly higher premium, but with the premium itself informed by the BMA probability that losses fall within the layer (computed by integrating the full posterior predictive mixture over the layer, not just evaluating it at a single point estimate). This final application illustrates the chapter's overarching message in its most concrete, decision-relevant form: the "right" reinsurance structure depends not on a single tail-risk number, but on the *full distribution* of plausible tail-risk numbers induced by parameter uncertainty (Section 20.1), prior elicitation (20.2), pooling structure (20.3), and dependence-model uncertainty (20.4) – precisely the synthesis that Bayesian tail inference, as developed across this Part, is designed to deliver.

Algorithm: end-to-end simulation of the integrated catastrophe model

Input: hierarchical posterior samples $\{(\xi_j^{(m)}, \sigma_j^{(m)})\}_{j=1}^4$ (from Algorithm 20.3); BMA model weights $\Pr(M_{\text{indep}}|\text{data}), \Pr(M_{\text{common}}|\text{data})$ (from Algorithm 20.4-type estimation); number of simulated "years" N.

1. For $n = 1, \dots, N$: draw a posterior sample index $m \sim \text{Uniform}\{1, \dots, M\}$ (giving $\{\xi_j^{(m)}, \sigma_j^{(m)}\}_{j=1}^4$ for this simulated year – propagating hierarchical posterior uncertainty).
2. Draw a dependence-model indicator $D_n \sim \text{Bernoulli}(\Pr(M_{\text{common}}|\text{data}))$.
3. If $D_n = 0$ (M_{indep}): for $j = 1, \dots, 4$, draw $L_{j,n} \sim \text{GPD}(\xi_j^{(m)}, \sigma_j^{(m)})$ independently.
4. If $D_n = 1$ (M_{common}): draw a common factor $Z_n \sim \text{LogNormal}(0, 0.3^2)$ (or the model's specified common-factor distribution), then for $j = 1, \dots, 4$, draw $L_{j,n} \sim \text{GPD}(\xi_j^{(m)}, \sigma_j^{(m)} \times Z_n)$ (the common factor scales each region's severity).
5. $\Lambda_{\text{agg},n} = \sum_{j=1}^4 L_{j,n}$.
6. After all N years: $\hat{\Lambda}^{\text{BMA}}_{0.996} =$ the 99.6% empirical quantile of $\{\Lambda_{\text{agg},n}\}_{n=1}^N$.

This algorithm is the computational synthesis of the entire chapter: step 1 propagates parameter uncertainty (Section 20.1, informed by hierarchical pooling, Section 20.3); steps 2–4 propagate dependence-model uncertainty (Section 20.4); and step 6's empirical quantile of N simulated years is itself subject to the rare-event-estimation considerations of Part 19 if N is not large enough relative to $1 - 0.996 = 0.004$ (a "1-in-250" probability needs $N \gg 250$, ideally $N \gtrsim 10^4$-10^5 for a stable quantile estimate per the diagnostic of Algorithm 20.1) – closing the loop back to the rare-event simulation methods of Part 19, now applied not to a single fixed model but to the full BMA-and-hierarchically-pooled predictive distribution

that this final Part has constructed.

Key formulas: Section 20.5

- One big jump: $\Pr(\Lambda_{\text{agg}} > x) \sim \sum_j \Pr(L_j > x)$ as $x \to \infty$ (independence); $\xi_{\text{agg}} \approx \max_j \xi_j$.
- Common factor: $L_j|\xi_j, \sigma_j, Z \sim \text{GPD}(\xi_j, \sigma_j Z)$, inducing positive dependence and heavier aggregate tails.
- BMA aggregate quantile $\neq$ probability-weighted average of component quantiles (mixture quantiles are shape-dependent); heaviest-tailed component dominates extreme quantiles of the mixture.

Notation and further reading. Catastrophe risk modelling, integrating the statistical (Sections 20.1–20.4), dependence-structure (this section), and physically-based simulation elements mentioned in Section 20.5's remarks, is a substantial applied field with its own dedicated literature – Mitchell-Wallace et al.'s *Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide* provides an industry-oriented overview, while the academic literature on multivariate extreme value theory and tail dependence (e.g. Beirlant et al., *Statistics of Extremes: Theory and Applications*) provides the rigorous statistical foundations for the dependence-structure questions raised in this closing section – and, with this section, brings Volume V's arc from abstract rate-function theory (Part 17) to applied catastrophe risk modelling to its conclusion.

Cross-volume connections. This closing case study's "one big jump vs. common factor" dichotomy for aggregate tail risk is the direct statistical-inference counterpart of Volume IV's "independent failures vs. cascading failures" dichotomy for network resilience – both ask whether system-level extreme outcomes are dominated by a single extreme component or by correlated/cascading extremity across many components, and both find that the correlated/cascading case typically dominates in realistic systems, with major practical (capital adequacy here; redundancy design in Volume IV) consequences. The volume's overall arc – from abstract rate-function theory (Part 17) through dynamic/metastable extensions (Part 18), computational rare-event methods (Part 19), to applied statistical inference (Part 20) – mirrors, at the level of mathematical methodology, this monograph series' broader progression from foundational mathematics (Mathematics of Robustness, Risk & Failure, Volumes I–VI) toward the applied network, biological, and computational themes developed across the series' companion volumes.

Appendices

Technical Background and Reference Material

Self-contained background on convex analysis, Perron–Frobenius theory, stochastic calculus, and Bayesian computation, together with full proofs of the central theorems of Part 17 and reference tables for the distributions used throughout the volume.

Example — Extended discussion: sensitivity to the common-factor distribution

Section 20.5 used LogNormal$(0, 0.3^2)$ for the common factor (mean ≈ 1.046, variance ≈ 0.103). A moment-matched Gamma(k, θ) has $k \approx 10.61, \theta \approx 0.0986$.

Tail comparison. At the 99.9% quantile: LogNormal gives $e^{0.3 \times 3.09} \approx 2.527$; Gamma (Wilson–Hilferty) gives ≈ 2.328 – LogNormal exceeds Gamma by $\approx 9\%$ *at matched mean and variance*, reflecting LogNormal's heavier tail.

Propagation. A 9% difference in the common factor's tail does not translate directly to

a 9% difference in Λ_{agg}'s quantile – it depends on how deep into the common factor's tail the 1-in-250-year aggregate event draws from. If "moderately elevated" (90–95th percentile, where LogNormal/Gamma are similar by moment-matching), little sensitivity; if "common factor itself extreme" (99th+), more sensitivity (though likely still $< 10\%$).

Recommendation. Re-running Algorithm 20.5 with a moment-matched alternative distribution is a cheap, informative sensitivity check: small sensitivity suggests the distributional *form* isn't critical (though its existence/strength, $\Pr(M_{\text{common}}|\text{data})$ and $\text{sd}(\log Z)$, remain so); large sensitivity warrants empirically estimating the common factor's shape from historical co-movement data – extending Section 20.3's pooling machinery to the dependence structure itself.

Example — Extended discussion: pricing a reinsurance layer from the aggregate loss distribution

A reinsurance layer $[d, d+l]$ pays $\min(\max(\Lambda_{\text{agg}}-d, 0), l)$. Its pure premium is $\mathbb{E}[(\Lambda_{\text{agg}}-d)^+] - \mathbb{E}[(\Lambda_{\text{agg}}-(d+l))^+]$ – a difference of two stop-loss transforms, each computable from Algorithm 20.5's simulated draws λ_i by averaging $(\lambda_i - d)^+$.

Tail sensitivity. For high attachments d (layers covering only extreme events), the pure premium is dominated by the same tail region where the common-factor distributional choice (this section's earlier extended discussion) matters most – a 9% difference in a 99.9% quantile can translate into a substantially larger relative difference in the pure premium of a layer attaching near that quantile, since the payoff is a convex-then-flat function of Λ_{agg}. This is why high-layer reinsurance pricing is acutely sensitive to tail assumptions that barely affect the median or 95th percentile of the same distribution.

Appendix A Convex Analysis and the Legendre–Fenchel Transform

Overview

This appendix collects the convex-analytic facts used throughout Part 17, centred on the Legendre–Fenchel transform $\Lambda^\star(x) = \sup_\lambda(\lambda x - \Lambda(\lambda))$ that converts a cumulant generating function into a large-deviation rate function. The results are stated for $\Lambda : \mathbb{R} \to (-\infty, +\infty]$, with the multivariate extension to $\mathbb{R}^d$ noted at the end.

Definition: Convex function, epigraph, lower semicontinuity

A function $\Lambda : \mathbb{R} \to (-\infty, +\infty]$ is *convex* if its epigraph $\{(\lambda, t) : t \geq \Lambda(\lambda)\}$ is a convex subset of $\mathbb{R}^2$, equivalently $\Lambda(\theta\lambda_1 + (1-\theta)\lambda_2) \leq \theta\Lambda(\lambda_1) + (1-\theta)\Lambda(\lambda_2)$ for $\theta \in [0,1]$. Λ is *lower semicontinuous* (*lsc*) if $\{\lambda : \Lambda(\lambda) \leq c\}$ is closed for every $c \in \mathbb{R}$, equivalently $\Lambda(\lambda_0) \leq \liminf_{\lambda\to\lambda_0} \Lambda(\lambda)$. The *domain* is $\mathcal{D}_\Lambda = \{\lambda : \Lambda(\lambda) < \infty\}$, always an interval (possibly unbounded, possibly all of $\mathbb{R}$) for convex Λ.

Why CGFs are convex and lsc. For $\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, convexity follows from Hölder's inequality: for $\theta \in [0,1]$ and conjugate exponents $p = 1/\theta, q = 1/(1-\theta)$,

$$\mathbb{E}[e^{(\theta\lambda_1+(1-\theta)\lambda_2)X}] = \mathbb{E}[(e^{\lambda_1 X})^\theta (e^{\lambda_2 X})^{1-\theta}] \leq \mathbb{E}[e^{\lambda_1 X}]^\theta \, \mathbb{E}[e^{\lambda_2 X}]^{1-\theta}$$

by Hölder, and taking logarithms gives $\Lambda(\theta\lambda_1+(1-\theta)\lambda_2) \leq \theta\Lambda(\lambda_1)+(1-\theta)\Lambda(\lambda_2)$, convexity. Lower semicontinuity follows from Fatou's lemma applied to $e^{\lambda X}$ as $\lambda \to \lambda_0$. Both properties

are thus automatic for any CGF, with no further assumptions on the underlying distribution ν beyond the existence of $\mathbb{E}[e^{\lambda X}]$ (possibly $= \infty$) for each λ.

Example — Theorem: Properties of the Legendre–Fenchel transform

Let $\Lambda : \mathbb{R} \to (-\infty, +\infty]$ be convex, lsc, with $\Lambda(0) = 0$ (the normalisation appropriate for CGFs, since $\mathbb{E}[e^{0 \cdot X}] = 1$). Then $\Lambda^\star(x) = \sup_\lambda(\lambda x - \Lambda(\lambda))$ satisfies:

(i) $\Lambda^\star$ is convex and lsc on $\mathbb{R}$ (as a pointwise supremum of affine functions $\lambda \mapsto \lambda x - \Lambda(\lambda)$, each affine in x for fixed λ).

(ii) $\Lambda^\star(x) \geq 0$ for all x (take $\lambda = 0$: $\Lambda^\star(x) \geq 0 \cdot x - \Lambda(0) = 0$).

(iii) **Biconjugacy:** $\Lambda^{\star\star} = \Lambda$, i.e. $\Lambda(\lambda) = \sup_x(\lambda x - \Lambda^\star(x))$ – the Legendre–Fenchel transform is an involution on convex lsc functions (the Fenchel–Moreau theorem).

(iv) If Λ is differentiable at λ_0 with $\Lambda'(\lambda_0) = x_0$, then $\Lambda^\star(x_0) = \lambda_0 x_0 - \Lambda(\lambda_0)$ and the supremum defining $\Lambda^\star(x_0)$ is attained at $\lambda = \lambda_0$ (first-order condition: $\frac{d}{d\lambda}(\lambda x_0 - \Lambda(\lambda)) = x_0 - \Lambda'(\lambda) = 0$ at $\lambda = \lambda_0$).

(v) If Λ is twice differentiable with $\Lambda'' > 0$ on $\mathcal{D}_\Lambda^\circ$ (strict convexity), then $\Lambda^\star$ is also twice differentiable on the corresponding range, with $(\Lambda^\star)''(x_0) = 1/\Lambda''(\lambda_0)$ for $x_0 = \Lambda'(\lambda_0)$ – the second derivatives of Λ and $\Lambda^\star$ are reciprocal at corresponding points, a relation sometimes called the "inverse function rule" for Legendre transforms.

Proof sketch of (iii), biconjugacy. The inequality $\Lambda^{\star\star}(\lambda) \leq \Lambda(\lambda)$ is immediate: for any x, $\lambda x - \Lambda^\star(x) \leq \lambda x - (\lambda x - \Lambda(\lambda)) = \Lambda(\lambda)$ (using $\Lambda^\star(x) \geq \lambda x - \Lambda(\lambda)$ by definition of the sup), so taking the sup over x on the left preserves the inequality. The reverse inequality $\Lambda^{\star\star}(\lambda) \geq \Lambda(\lambda)$ is the substantive part, and requires Λ to be convex and lsc: it follows from the *supporting hyperplane theorem* – the epigraph of a convex lsc function is the intersection of all the closed half-spaces (supporting hyperplanes) that contain it, and each supporting hyperplane with slope λ touching the epigraph at height $\Lambda(\lambda)$ corresponds to a point x at which $\lambda x - \Lambda^\star(x) = \Lambda(\lambda)$ – so the supremum defining $\Lambda^{\star\star}(\lambda)$ recovers $\Lambda(\lambda)$ exactly. This is the convex-analytic content underlying the "rate function determines the CGF and vice versa" duality used throughout Section 17.2.

Worked illustration of property (v). For the Gaussian CGF $\Lambda(\lambda) = \mu\lambda + \frac{1}{2}\sigma^2\lambda^2$, $\Lambda''(\lambda) = \sigma^2$ (constant), so $(\Lambda^\star)''(x) = 1/\sigma^2$ (also constant) – consistent with $\Lambda^\star(x) = (x-\mu)^2/(2\sigma^2)$, whose second derivative is indeed $1/\sigma^2$. For the Exponential CGF $\Lambda(\lambda) = -\log(1 - \lambda/\theta)$ $(\theta > 0, \lambda < \theta)$, $\Lambda'(\lambda) = 1/(\theta - \lambda)$, so $\Lambda''(\lambda) = 1/(\theta-\lambda)^2$. At λ_0 with $x_0 = \Lambda'(\lambda_0) = 1/(\theta - \lambda_0)$, i.e. $\theta - \lambda_0 = 1/x_0$, we get $\Lambda''(\lambda_0) = x_0^2$, so $(\Lambda^\star)''(x_0) = 1/x_0^2$. Direct differentiation of $\Lambda^\star(x) = \theta x - 1 - \log(\theta x)$ (Section 17.2) gives $(\Lambda^\star)'(x) = \theta - 1/x$ and $(\Lambda^\star)''(x) = 1/x^2$ – matching exactly.

Multivariate extension

For $\Lambda : \mathbb{R}^d \to (-\infty, +\infty]$ convex and lsc with $\Lambda(\mathbf{0}) = 0$, the Legendre–Fenchel transform $\Lambda^\star(\mathbf{x}) = \sup_{\square \in \mathbb{R}^d}(\square \cdot \mathbf{x} - \Lambda(\square))$ retains properties (i)–(iv) verbatim, with the differentiability condition (iv) becoming $\nabla\Lambda(\square_0) = \mathbf{x}_0$, and property (v) becoming a matrix relation: if $\nabla^2\Lambda(\square_0) \succ 0$ (positive definite Hessian), then $\nabla^2\Lambda^\star(\mathbf{x}_0) = [\nabla^2\Lambda(\square_0)]^{-1}$ – the Hessians of Λ and $\Lambda^\star$ at corresponding points are matrix inverses of each other. This matrix relation is the multivariate analogue used implicitly whenever a multivariate Gaussian rate function $\Lambda^\star(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \square)^\top \square^{-1}(\mathbf{x} - \square)$ (covariance $\square$) is paired with its CGF $\Lambda(\square) = \square \cdot \square + \frac{1}{2}\square^\top \square\square$ (note: $\square$, not $\square^{-1}$, appears in Λ) – the Hessians $\square^{-1}$

and $\square$ are indeed inverses, exactly as property (v) requires.

Sub-differentials and non-differentiable Λ. When Λ is convex but not differentiable at some λ_0 (a "kink"), the *sub-differential* $\partial\Lambda(\lambda_0) = \{x : \Lambda(\lambda) \geq \Lambda(\lambda_0) + x(\lambda - \lambda_0)\ \forall\lambda\}$ is a non-degenerate closed interval $[x^-, x^+]$ (rather than a single point), and property (iv) generalises: $\Lambda^\star(x) = \lambda_0 x - \Lambda(\lambda_0)$ for *all* $x \in \partial\Lambda(\lambda_0)$ – i.e. $\Lambda^\star$ is *affine* (linear) on the interval $[x^-, x^+]$, with slope λ_0. Dually, if $\Lambda^\star$ has a kink at x_0, Λ is affine on the corresponding sub-differential interval of $\Lambda^\star$ at x_0. This duality between "kinks in one function" and "affine segments in the other" is the precise mechanism underlying the non-steep CGF phenomenon discussed in Section 17.4's remarks: a CGF Λ that approaches a finite slope $L = \lim_{\lambda\to\lambda_{\max}} \Lambda'(\lambda)$ as $\lambda \to \lambda_{\max}^-$ (the boundary of its domain) corresponds to $\Lambda^\star$ becoming affine (with slope $\lambda_{\max}$) for all $x \geq L$ – not literally a "kink" in Λ itself, but a boundary effect with the same qualitative consequence (loss of strict convexity of $\Lambda^\star$ beyond $x = L$).

Example — Worked example: the indicator-function / hard-constraint limit

Consider $\Lambda_M(\lambda) = M|\lambda|$ for large $M > 0$ (a CGF-like function with a kink at $\lambda = 0$, where the left- and right-derivatives are $-M$ and $+M$ respectively, so $\partial\Lambda_M(0) = [-M, M]$). Its Legendre transform: $\Lambda_M^\star(x) = \sup_\lambda(\lambda x - M|\lambda|)$. For $|x| \leq M$, the supremum is at $\lambda = 0$ (any $\lambda > 0$ gives $\lambda(x - M) \leq 0$ for $x \leq M$, and any $\lambda < 0$ gives $\lambda(x + M) \leq 0$ for $x \geq -M$), so $\Lambda_M^\star(x) = 0$ for $|x| \leq M$. For $|x| > M$, the supremum is $+\infty$ (taking $\lambda \to \pm\infty$ with sign matching x). So $\Lambda_M^\star(x) = 0$ for $x \in [-M, M]$ and $+\infty$ otherwise – as $M \to \infty$, $\Lambda_M \to 0$ everywhere (the "trivial" CGF of a point mass at 0) while $\Lambda_\infty^\star \equiv 0$ on $\mathbb{R}$ – consistent with biconjugacy: the CGF of a point mass at 0 is $\Lambda \equiv 0$, whose Legendre transform is $\Lambda^\star \equiv 0$ (the rate function of a deterministic sequence is identically zero, since there is no "deviation" to be rare). This degenerate example illustrates the boundary behaviour of properties (i)–(v) when Λ itself becomes degenerate, a useful sanity check when implementing Algorithm 17.2's Newton iteration (which should detect $\Lambda'' \to 0$ and handle the resulting ill-conditioning gracefully, e.g. by falling back to a bisection step).

Appendix B Perron–Frobenius Theory of Non-negative Matrices

Overview

The Perron–Frobenius theorem underlies the Gärtner–Ellis rate function for Markov-chain additive functionals (Section 17.4) and the Donsker–Varadhan formula (Section 18.3): both reduce to computing $\rho(\mathbf{P}_\lambda)$, the spectral radius (dominant eigenvalue) of a non-negative "tilted" matrix. This appendix states the theorem and the properties of $\rho(\mathbf{P}_\lambda)$ as a function of λ that make $\Lambda(\lambda) = \log\rho(\mathbf{P}_\lambda)$ a valid CGF.

Definition: Irreducible non-negative matrices

A matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ with $A_{ij} \geq 0$ for all i, j is *irreducible* if for every pair (i, j) there exists $n \geq 1$ such that $(\mathbf{A}^n)_{ij} > 0$ – equivalently, the directed graph with an edge $i \to j$ whenever $A_{ij} > 0$ is strongly connected (every node reachable from every other node). $\mathbf{A}$ is *primitive* (or aperiodic) if additionally there exists a single n_0 such that $(\mathbf{A}^{n_0})_{ij} > 0$ for *all* (i, j) simultaneously (the graph has no cyclic periodicity).

Example — Theorem: Perron–Frobenius (irreducible non-negative matrices)

Let $\mathbf{A} \geq 0$ be irreducible, $K \times K$. Then:

(i) $\rho(\mathbf{A}) > 0$ is a (real, positive) eigenvalue of $\mathbf{A}$ – the *Perron root* – and is a *simple* eigenvalue (algebraic multiplicity 1).

(ii) There exists a corresponding eigenvector $\mathbf{v} > 0$ (all entries strictly positive), unique up to positive scaling – the *Perron eigenvector*.

(iii) $\rho(\mathbf{A}) > |\mu|$ for every other eigenvalue μ of $\mathbf{A}$, *unless* $\mathbf{A}$ is periodic with period $d > 1$, in which case there are exactly d eigenvalues of modulus $\rho(\mathbf{A})$ (the d-th roots of $\rho(\mathbf{A})^d$ times $\rho(\mathbf{A})$, i.e. $\rho(\mathbf{A})e^{2\pi ik/d}$ for $k = 0, \dots, d-1$). For primitive $\mathbf{A}$, $\rho(\mathbf{A})$ is the *unique* eigenvalue of maximum modulus.

(iv) **Collatz–Wielandt variational formula:** $\rho(\mathbf{A}) = \max_{\mathbf{v}>0} \min_i \frac{(\mathbf{Av})_i}{v_i} = \min_{\mathbf{v}>0} \max_i \frac{(\mathbf{Av})_i}{v_i}$.

Why irreducibility matters for $\mathbf{P}_\lambda$. The tilted matrix $(\mathbf{P}_\lambda)_{ij} = P_{ij}e^{\lambda f(j)}$ has exactly the same zero/non-zero pattern as $\mathbf{P}$ (since $e^{\lambda f(j)} > 0$ always), so $\mathbf{P}_\lambda$ is irreducible whenever $\mathbf{P}$ is – a property of the underlying chain's transition structure, unaffected by the tilt. This guarantees, via (i)–(ii), that $\rho(\mathbf{P}_\lambda)$ is a well-defined, simple, positive eigenvalue with a strictly positive eigenvector for *every* $\lambda \in \mathbb{R}$ – exactly the regularity needed for $\Lambda(\lambda) = \log \rho(\mathbf{P}_\lambda)$ to be a smooth function of λ (simple eigenvalues of a matrix depend analytically on smooth perturbations of the matrix, by standard analytic perturbation theory – a simple eigenvalue cannot "collide" with another eigenvalue and lose smoothness, unlike a repeated eigenvalue, which can exhibit non-smooth, e.g. square-root-type, dependence on parameters).

Theorem: Convexity of $\lambda \mapsto \log \rho(\mathbf{P}_\lambda)$

For $(\mathbf{P}_\lambda)_{ij} = P_{ij}e^{\lambda f(j)}$, write $\rho(\mathbf{P}_\lambda) = \max_{\mathbf{v}>0} \min_i \frac{\sum_j P_{ij}e^{\lambda f(j)}v_j}{v_i}$ (Collatz–Wielandt, property (iv) above). For *fixed* $\mathbf{v} > 0$, the function $\lambda \mapsto \log\left(\frac{\sum_j P_{ij}e^{\lambda f(j)}v_j}{v_i}\right)$ is convex in λ for each i (it is the log of a sum of exponentials $e^{\lambda f(j)} \times$ (positive constant), and $\log \sum_k e^{\lambda c_k + d_k}$ is convex in λ for any constants c_k, d_k – a standard "log-sum-exp is convex" fact, itself a consequence of Hölder's inequality as in Appendix A). The rigorous proof that $\lambda \mapsto \log \rho(\mathbf{P}_\lambda)$ itself is convex proceeds via a sub-multiplicativity argument on $\mathbf{P}_\lambda$'s entries combined with continuity of the spectral radius, rather than the variational formula directly – but the variational formula provides useful intuition for *why* convexity should hold (each "candidate" ρ-value, for fixed $\mathbf{v}$, is itself log-convex in λ).

Numerical conditioning of power iteration (Algorithm 17.4). The convergence rate of power iteration is governed by $|\mu_2(\mathbf{P}_\lambda)|/\rho(\mathbf{P}_\lambda)$, the ratio of the second-largest eigenvalue modulus to the Perron root – the *spectral gap*. For finite-state chains with bounded f, $\mathcal{D}_\Lambda = \mathbb{R}$ always, but the spectral gap can still shrink for large $|\lambda|$, slowing power iteration – a practical reason to prefer, for large $|\lambda|$, either (a) shift-invert variants of power iteration, or (b) the eigenvector-perturbation formula for $\Lambda'(\lambda)$ noted in Algorithm 17.4, which avoids rerunning power iteration from scratch at each λ by tracking how the eigenvector itself changes smoothly with λ (a "continuation" approach, in the same spirit as Algorithm 17.5's continuation strategy for contraction-principle optimisations).

Example — Worked example: spectral gap for the regime-switching chain

For the chain $\mathbf{P} = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$ (Section 17.4's $\alpha = \beta = 0.3$ example), the eigenvalues are found from $\det(\mathbf{P} - \mu\mathbf{I}) = 0$: $(0.7-\mu)^2 - 0.09 = 0 \Rightarrow 0.7 - \mu = \pm 0.3 \Rightarrow \mu \in \{1, 0.4\}$. So $\rho(\mathbf{P}) = 1$ (as required, $\Lambda(0) = 0$) and $\mu_2 = 0.4$, giving a spectral gap of $1 - 0.4 = 0.6$ – a moderately fast-mixing chain (mixing time $\sim 1/(1-\mu_2) \approx 1.67$ steps). For a more persistent chain, e.g. $\alpha = \beta = 0.05$ (each state has 95% probability of self-transition), $\mathbf{P} = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}$, eigenvalues $\{1, 0.9\}$, spectral gap 0.1, mixing time ~ 10 steps – power iteration for $\rho(\mathbf{P}_\lambda)$ on this more persistent chain converges roughly $6\times$ more slowly (per the $|\mu_2|/\rho$ convergence-rate scaling) than for the $\alpha = \beta = 0.3$ chain, a concrete illustration of how a chain's physical persistence (relevant, e.g., to the regime-switching volatility application of Section 17.4) directly impacts the computational cost of evaluating its large-deviation rate function.

Appendix C Full Proofs: Cramér's Theorem and Sanov's Theorem

Overview

Sections 17.2 and 17.3 presented proof *sketches* of Cramér's and Sanov's theorems, focused on conveying the key ideas (Chernoff bounds, exponential tilting, the method of types) without full measure-theoretic rigour. This appendix fills in the technical steps elided there, for the reader wishing to see the arguments in complete detail.

Example — Theorem: Cramér's theorem, full statement and proof

Setup. $X_1, X_2, \ldots$ i.i.d. on $\mathbb{R}$ with law ν, $\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda X}] < \infty$ for λ in an open neighbourhood of 0 (ensuring Λ is finite, smooth, and strictly convex near 0, by standard properties of moment generating functions). **Claim:** $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies the LDP with rate n and good convex rate function $I = \Lambda^\star$.

Upper bound (full detail). For any closed set $C \subseteq \mathbb{R}$ and any $\lambda \in \mathbb{R}$, by Markov's inequality applied to the non-negative random variable $e^{\lambda S_n}$ (where $S_n = n\bar{X}_n$): fix $x \in C$ and λ with the same sign as $x - \mathbb{E}[X]$; then $\Pr(\bar{X}_n \geq x) \leq e^{-n(\lambda x - \Lambda(\lambda))}$ for $\lambda \geq 0$ (the computation of Section 17.2's proof sketch), and similarly $\Pr(\bar{X}_n \leq x) \leq e^{-n(\lambda x - \Lambda(\lambda))}$ for $\lambda \leq 0$. Taking the infimum over λ of the appropriate sign gives $\Pr(\bar{X}_n \geq x) \leq e^{-n \sup_{\lambda \geq 0}(\lambda x - \Lambda(\lambda))}$. For $x > \mathbb{E}[X] = \Lambda'(0)$, the unconstrained supremum $\sup_{\lambda \in \mathbb{R}}(\lambda x - \Lambda(\lambda)) = \Lambda^\star(x)$ is attained at some $\lambda^\star(x) > 0$ (since $\Lambda'(\lambda^\star) = x > \Lambda'(0)$ and Λ' is increasing by convexity), so the constraint $\lambda \geq 0$ is not binding and $\sup_{\lambda \geq 0}(\lambda x - \Lambda(\lambda)) = \Lambda^\star(x)$. Hence $\Pr(\bar{X}_n \geq x) \leq e^{-n\Lambda^\star(x)}$ for $x > \mathbb{E}[X]$, and symmetrically for $x < \mathbb{E}[X]$. For a general closed C not containing $\mathbb{E}[X]$, $C \subseteq (-\infty, \sup C] \cup [\inf C, \infty)$ and the bound applies to whichever half-line is relevant; for C containing $\mathbb{E}[X]$ in its interior, $\inf_C I = 0 = I(\mathbb{E}[X])$ trivially (since $\Pr(\bar{X}_n \in C) \leq 1 = e^{-n \times 0}$). This establishes $\limsup_n \frac{1}{n} \log \Pr(\bar{X}_n \in C) \leq -\inf_{x \in C} \Lambda^\star(x)$ for all closed C.

Lower bound (full detail, via tilting). Let U be open, $x \in U$, and suppose Λ is differentiable at the corresponding $\lambda^\star = \lambda^\star(x)$ with $\Lambda'(\lambda^\star) = x$ (true for x in the interior of $\Lambda'(\mathcal{D}_\Lambda^\circ)$, the "steep" region). Define the tilted measure $\nu_{\lambda^\star}(dy) = e^{\lambda^\star y - \Lambda(\lambda^\star)}\nu(dy)$, under which $X_1, \ldots, X_n$ i.i.d. $\sim \nu_{\lambda^\star}$ have mean $\mathbb{E}_{\nu_{\lambda^\star}}[X] = \Lambda'(\lambda^\star) = x$. By the weak law of large numbers under $\nu_{\lambda^\star}$, $\Pr_{\nu_{\lambda^\star}}(\bar{X}_n \in U) \to 1$ as $n \to \infty$ (since $x \in U$ open

and $\bar{X}_n \to x$ in $\nu_{\lambda^\star}$-probability). Now relate $\Pr_\nu(\bar{X}_n \in U)$ to $\Pr_{\nu_{\lambda^\star}}(\bar{X}_n \in U)$ via the Radon–Nikodym derivative: for i.i.d. sequences, $\frac{d\nu^{\otimes n}}{d\nu_{\lambda^\star}^{\otimes n}}(x_1, \dots, x_n) = \prod_{i=1}^n \frac{d\nu}{d\nu_{\lambda^\star}}(x_i) = \prod_i e^{-\lambda^\star x_i + \Lambda(\lambda^\star)} = e^{-\lambda^\star S_n + n\Lambda(\lambda^\star)} = e^{-n(\lambda^\star \bar{X}_n - \Lambda(\lambda^\star))}$. So

$$\Pr_\nu(\bar{X}_n \in U) = \mathbb{E}_{\nu_{\lambda^\star}}\left[e^{-n(\lambda^\star \bar{X}_n - \Lambda(\lambda^\star))} \mathbf{1}_{\bar{X}_n \in U}\right].$$

On the event $\{\bar{X}_n \in U\} \cap \{|\bar{X}_n - x| < \delta\}$ (for small δ, an event whose $\nu_{\lambda^\star}$-probability also $\to 1$ since $\bar{X}_n \to x$ in $\nu_{\lambda^\star}$-probability and $x \in U$ open), $\lambda^\star \bar{X}_n - \Lambda(\lambda^\star) \in (\Lambda^\star(x) - |\lambda^\star|\delta, \Lambda^\star(x) + |\lambda^\star|\delta)$ (using property (iv) of Appendix A, $\Lambda^\star(x) = \lambda^\star x - \Lambda(\lambda^\star)$). So $e^{-n(\lambda^\star \bar{X}_n - \Lambda(\lambda^\star))} \geq e^{-n(\Lambda^\star(x) + |\lambda^\star|\delta)}$ on this event, giving

$$\Pr_\nu(\bar{X}_n \in U) \geq e^{-n(\Lambda^\star(x)+|\lambda^\star|\delta)} \Pr_{\nu_{\lambda^\star}}(\bar{X}_n \in U,\ |\bar{X}_n - x| < \delta) \to e^{-n(\Lambda^\star(x)+|\lambda^\star|\delta)} \times 1$$

as $n \to \infty$ (the $\nu_{\lambda^\star}$-probability $\to 1$). Hence $\liminf_n \frac{1}{n} \log \Pr_\nu(\bar{X}_n \in U) \geq -\Lambda^\star(x) - |\lambda^\star|\delta$. Letting $\delta \to 0$ gives $\liminf_n \frac{1}{n} \log \Pr_\nu(\bar{X}_n \in U) \geq -\Lambda^\star(x)$. Taking $x \in U$ to approach $\inf_{x \in U} \Lambda^\star(x)$ gives $\liminf_n \frac{1}{n} \log \Pr_\nu(\bar{X}_n \in U) \geq -\inf_{x \in U} \Lambda^\star(x)$, the LDP lower bound for open sets.

Goodness of $I = \Lambda^\star$. $\Lambda^\star$ is good (compact sub-level sets) because $\Lambda^\star(x) \to \infty$ as $|x| \to \infty$ whenever Λ is finite in a neighbourhood of 0 (the supremum $\sup_\lambda(\lambda x - \Lambda(\lambda)) \geq \lambda_0 x - \Lambda(\lambda_0) \to \infty$ as $x \to \pm\infty$ for any fixed $\lambda_0 \neq 0$ with $\Lambda(\lambda_0) < \infty$), combined with lower semicontinuity (Appendix A) – sub-level sets $\{\Lambda^\star \leq c\}$ are then closed and bounded, hence compact in $\mathbb{R}$.

Example — Theorem: Sanov's theorem, full statement and proof (finite alphabet)

Setup. $X_1, \dots, X_n$ i.i.d. on finite alphabet $\mathcal{A} = \{1, \dots, K\}$ with law $\nu = (\nu_1, \dots, \nu_K)$, all $\nu_j > 0$. $\hat{\nu}_n = \frac{1}{n}\sum_i \delta_{X_i} \in \Delta_K$ (the probability simplex). **Claim:** $(\hat{\nu}_n)$ satisfies the LDP on Δ_K (a finite-dimensional compact set) with rate n and good convex rate function $I(\rho) = D(\rho\|\nu)$.

Exact probability via types. For $\rho \in \Delta_K$ such that $n\rho_j \in \mathbb{Z}_{\geq 0}$ for all j (an "achievable type" for sample size n), $\Pr(\hat{\nu}_n = \rho) = \binom{n}{n\rho_1, \dots, n\rho_K} \prod_{j=1}^K \nu_j^{n\rho_j}$ exactly (multinomial probability).

Stirling bounds on the multinomial coefficient. Using $\sqrt{2\pi m}\, m^m e^{-m} \leq m! \leq e\sqrt{m}\, m^m e^{-m}$ (a standard two-sided Stirling bound, valid for all $m \geq 1$), one obtains, after some algebra, the two-sided bound

$$\frac{1}{(n+1)^K} e^{nH(\rho)} \leq \binom{n}{n\rho_1, \dots, n\rho_K} \leq e^{nH(\rho)}$$

where $H(\rho) = -\sum_j \rho_j \log \rho_j$ is the Shannon entropy (the upper bound follows from $\binom{n}{n\rho_1, \dots, n\rho_K} \leq \frac{n!}{\prod_j (n\rho_j)!}$ combined with the Stirling *lower* bounds on each $(n\rho_j)!$ in the denominator and *upper* bound on $n!$ in the numerator; the lower bound similarly with the roles reversed, and the $(n+1)^K$ factor arising because there are at most $(n+1)^K$ distinct achievable types, so the total probability $\sum_{\text{types}} \Pr(\hat{\nu}_n = \rho) = 1$ combined with each term being within a factor $(n+1)^K$ of $e^{n[H(\rho) + \sum_j \rho_j \log \nu_j]}$ pins down the normalisation).

Combining. $\Pr(\hat{\nu}_n = \rho) = \binom{n}{n\rho_1, \ldots, n\rho_K} \prod_j \nu_j^{n\rho_j}$, and $\log \prod_j \nu_j^{n\rho_j} = n \sum_j \rho_j \log \nu_j$. So

$$\frac{1}{(n+1)^K} e^{n[H(\rho) + \sum_j \rho_j \log \nu_j]} \leq \Pr(\hat{\nu}_n = \rho) \leq e^{n[H(\rho) + \sum_j \rho_j \log \nu_j]},$$

and $H(\rho) + \sum_j \rho_j \log \nu_j = -\sum_j \rho_j \log \rho_j + \sum_j \rho_j \log \nu_j = -\sum_j \rho_j \log \frac{\rho_j}{\nu_j} = -D(\rho\|\nu)$. So

$$\frac{1}{(n+1)^K} e^{-nD(\rho\|\nu)} \leq \Pr(\hat{\nu}_n = \rho) \leq e^{-nD(\rho\|\nu)},$$

giving $\frac{1}{n} \log \Pr(\hat{\nu}_n = \rho) \to -D(\rho\|\nu)$ exactly (the $(n+1)^K$ factor contributes $\frac{K \log(n+1)}{n} \to 0$).

From points to sets. For a set $A \subseteq \Delta_K$, $\Pr(\hat{\nu}_n \in A) = \sum_{\rho \in A \cap \mathcal{T}_n} \Pr(\hat{\nu}_n = \rho)$ where $\mathcal{T}_n$ is the (finite, $\leq (n+1)^K$-element) set of achievable types. The upper bound $\Pr(\hat{\nu}_n \in A) \leq |A \cap \mathcal{T}_n| \times \max_{\rho \in A} e^{-nD(\rho\|\nu)} \leq (n+1)^K e^{-n \inf_{\rho \in A} D(\rho\|\nu)}$ gives $\limsup_n \frac{1}{n} \log \Pr(\hat{\nu}_n \in A) \leq -\inf_{\bar{A}} D(\rho\|\nu)$ (for closed A, using continuity of $D(\cdot\|\nu)$ to relate $\inf_A$ and $\inf_{\bar{A}}$, with the polynomial $(n+1)^K$ factor vanishing in the rate). The lower bound for open A follows by selecting a single achievable type $\rho_n \in A \cap \mathcal{T}_n$ with $\rho_n \to \rho^\star = \arg\inf_A D(\cdot\|\nu)$ (possible since achievable types become dense in Δ_K) and using $\Pr(\hat{\nu}_n \in A) \geq \Pr(\hat{\nu}_n = \rho_n) \geq \frac{1}{(n+1)^K} e^{-nD(\rho_n\|\nu)} \to e^{-nD(\rho^\star\|\nu)}$ to leading exponential order, by continuity of $D(\cdot\|\nu)$.

Goodness. Δ_K is compact (closed and bounded in $\mathbb{R}^K$), so every closed subset of Δ_K is compact automatically, and $D(\cdot\|\nu)$ is continuous on all of Δ_K given $\nu_j > 0$ for all j (since $\rho_j \log(\rho_j/\nu_j) \to 0$ as $\rho_j \to 0$, by the convention $0 \log 0 = 0$), making $D(\cdot\|\nu)$ a good rate function on Δ_K trivially.

Appendix D Stochastic Calculus Background: Itô's Formula and Girsanov's Theorem

Overview

Part 18's Freidlin–Wentzell theory and the path-space tilting of Section 19.2 rest on two pillars of stochastic calculus: Itô's formula (for manipulating functions of diffusion processes) and Girsanov's theorem (for changing the drift of a diffusion via an absolutely continuous change of measure – the rigorous version of the "exponential tilting" of Section 19.2 lifted to path space). This appendix states both results and connects them explicitly to the action functional of Section 18.1.

Definition: Itô process and Itô's formula

An *Itô process* is $X_t = X_0 + \int_0^t b(X_s)\, ds + \int_0^t \sigma(X_s)\, dW_s$ (informally $dX_t = b(X_t)dt + \sigma(X_t)dW_t$), where W_t is standard Brownian motion and the second integral is an Itô stochastic integral. For $g \in C^2(\mathbb{R})$, *Itô's formula* states

$$dg(X_t) = g'(X_t)\, dX_t + \tfrac{1}{2} g''(X_t)\, \sigma(X_t)^2\, dt = [g'(X_t) b(X_t) + \tfrac{1}{2} g''(X_t) \sigma(X_t)^2]\, dt + g'(X_t)\sigma(X_t)\, dW_t.$$

The extra term $\frac{1}{2} g'' \sigma^2 dt$ (absent in ordinary calculus) arises because $(dW_t)^2 = dt$ "to leading order" (the quadratic variation of Brownian motion), a fact that can be made rigorous via the definition of the Itô integral as a limit of Riemann–Stieltjes-type sums.

Why Itô's formula matters for Part 18. The Hamilton–Jacobi equation for the quasi-potential V (Section 18.2) is derived by applying Itô's formula to $V(X_t^\epsilon)$ and demanding that $\mathbb{E}[dV(X_t^\epsilon)]/dt \to 0$ in an appropriate sense as $\epsilon \to 0$ (a "no drift in the WKB phase" condition) – the extra Itô term $\frac{1}{2}V''\sigma^2$ is precisely what produces the quadratic-in-∇V term $\frac{1}{2}|\nabla V|^2_{a^{-1}}$ in the HJ equation, after the WKB ansatz $p_\epsilon \sim e^{-V/\epsilon}$ is substituted into the (Itô-formula-derived) backward Kolmogorov equation $\partial_t u + b\cdot\nabla u + \frac{\epsilon}{2}\Delta_a u = 0$ – substituting $u = e^{-V/\epsilon}$ and collecting the leading-order (ϵ^0) terms as $\epsilon \to 0$ gives exactly $-V_t - b\cdot\nabla V + \frac{1}{2}|\nabla V|^2_a = 0$ (the steady-state version, with a^{-1} in place of a, matches Section 18.2's HJ equation via the corresponding forward equation for the invariant density). This WKB derivation is the rigorous bridge alluded to in Section 18.2's remarks.

Example — Theorem: Girsanov's theorem (constant-drift case)

Let W_t be standard Brownian motion under measure Pr, and let $\tilde{W}_t = W_t - \mu t$ for a constant $\mu \in \mathbb{R}^d$. Define $\widetilde{\Pr}$ via the Radon–Nikodym derivative (on the path space $C([0,T];\mathbb{R}^d)$)

$$\left.\frac{d\widetilde{\Pr}}{d\Pr}\right|_{\mathcal{F}_T} = \exp\left(\mu\cdot W_T - \tfrac{1}{2}|\mu|^2 T\right).$$

Then under $\widetilde{\Pr}$, $\tilde{W}_t$ is a standard Brownian motion – equivalently, under $\widetilde{\Pr}$, W_t has drift μ (i.e. $W_t = \tilde{W}_t + \mu t$ with $\tilde{W}$ a $\widetilde{\Pr}$-Brownian motion). More generally, for a (sufficiently regular, e.g. bounded and predictable) drift process u_t, the Radon–Nikodym derivative $\frac{d\widetilde{\Pr}}{d\Pr} = \exp\left(\int_0^T u_s\cdot dW_s - \frac{1}{2}\int_0^T |u_s|^2 ds\right)$ (the *Doléans-Dade exponential*) defines a measure $\widetilde{\Pr}$ under which $W_t - \int_0^t u_s\,ds$ is a Brownian motion.

Connection to the action functional (Section 18.1). For the small-noise SDE $dX_t^\epsilon = b(X_t^\epsilon)dt + \sqrt{\epsilon}\,\sigma(X_t^\epsilon)dW_t$, write $W_t = \frac{1}{\sqrt{\epsilon}}\int_0^t \sigma(X_s^\epsilon)^{-1}(dX_s^\epsilon - b(X_s^\epsilon)ds)$ (formally solving for the driving Brownian motion). To make a target path ϕ "typical," apply Girsanov with drift $u_t = \frac{1}{\sqrt{\epsilon}}\sigma(\phi(t))^{-1}(\dot\phi(t) - b(\phi(t)))$ – exactly the "control" $u(t) = \sigma^{-1}(\dot\phi - b(\phi))$ identified in Section 18.1's remarks. The Girsanov exponent is then

$$\int_0^T u_s\cdot dW_s - \frac{1}{2}\int_0^T |u_s|^2 ds, \qquad \text{with } \frac{1}{2}\int_0^T |u_s|^2 ds = \frac{1}{2\epsilon}\int_0^T (\dot\phi - b(\phi))^\top a(\phi)^{-1}(\dot\phi - b(\phi))\,dt = \frac{I_{[0,T]}(\phi)}{\epsilon}.$$

So the Radon–Nikodym derivative's "rate" term is exactly $-I_{[0,T]}(\phi)/\epsilon$ – the negative exponential rate appearing in the Freidlin–Wentzell LDP. This is the rigorous content behind the informal statement (Section 18.1) that "the action functional is the Girsanov cost of the tilt that makes ϕ typical": Girsanov's theorem provides the *exact* Radon–Nikodym derivative relating the original and tilted path measures, and the Freidlin–Wentzell LDP follows from analysing this Radon–Nikodym derivative as $\epsilon \to 0$, in close analogy to (and as an infinite-dimensional generalisation of) the finite-dimensional tilting argument of Appendix C's proof of Cramér's theorem.

Example — Worked example: Girsanov cost for the Schilder straight-line path

For $b \equiv 0, \sigma = \mathbf{I}_d$ (Schilder's theorem, Section 18.1), and target path $\phi(t) = \mathbf{y}\,t/T$, the control is $u(t) = \frac{1}{\sqrt{\epsilon}}\dot\phi(t) = \frac{\mathbf{y}}{\sqrt{\epsilon}\,T}$ (constant). The Girsanov exponent's quadratic term is $\frac{1}{2}\int_0^T |u(t)|^2 dt = \frac{1}{2}\cdot\frac{|\mathbf{y}|^2}{\epsilon T^2}\cdot T = \frac{|\mathbf{y}|^2}{2\epsilon T}$ – matching $I_{[0,T]}(\phi)/\epsilon = \frac{|\mathbf{y}|^2}{2T\epsilon}$ exactly (Section 18.1's exercises). The linear term $\int_0^T u_s\cdot dW_s = \frac{\mathbf{y}}{\sqrt{\epsilon}T}\cdot W_T$ is, under the *original* measure, a mean-zero Gaussian with variance $\frac{|\mathbf{y}|^2}{\epsilon T^2}\times T = \frac{|\mathbf{y}|^2}{\epsilon T}$ – so the full Radon–Nikodym

derivative $\exp(\int u\,dW - \frac{1}{2}\int |u|^2 ds)$ has $\mathbb{E}[\frac{d\widetilde{\Pr}}{d\Pr}] = 1$ (as required for any valid change of measure) despite the exponent's quadratic term alone being deterministic and equal to $|\mathbf{y}|^2/(2\epsilon T)$ – the random linear term exactly compensates in expectation, an instance of $\mathbb{E}[e^{Z-\frac{1}{2}\operatorname{Var}(Z)}] = 1$ for Gaussian Z with mean 0, here $Z = \int u\,dW \sim \mathcal{N}(0, \int |u|^2 ds)$.

Appendix E Markov Chain Monte Carlo: Metropolis–Hastings and Hamiltonian Monte Carlo

Overview

Part 20's Bayesian computations (Algorithms 20.1, 20.3, 20.4) all rely on MCMC to sample from posterior distributions $\pi(\theta \mid \text{data})$ known only up to a normalising constant. This appendix reviews the Metropolis–Hastings algorithm (the general-purpose workhorse) and Hamiltonian Monte Carlo (HMC, the modern default for continuous, differentiable posteriors such as the GPD likelihoods of Part 20), with attention to the convergence diagnostics needed to trust MCMC output.

Definition: Metropolis–Hastings algorithm

To sample from $\pi(\theta) \propto \tilde{\pi}(\theta)$ (with $\tilde{\pi}$ computable, π's normalising constant unknown): given current state $\theta^{(t)}$, propose $\theta' \sim q(\cdot \mid \theta^{(t)})$ (a *proposal distribution*), and accept $(\theta^{(t+1)} = \theta')$ with probability

$$\alpha(\theta^{(t)}, \theta') = \min\left(1,\ \frac{\tilde{\pi}(\theta')\,q(\theta^{(t)} \mid \theta')}{\tilde{\pi}(\theta^{(t)})\,q(\theta' \mid \theta^{(t)})}\right),$$

otherwise reject $(\theta^{(t+1)} = \theta^{(t)})$. For symmetric proposals $(q(\theta'|\theta) = q(\theta|\theta')$, e.g. a Gaussian random walk $\theta' = \theta^{(t)} + \mathcal{N}(0, \Sigma_{\text{prop}}))$, the proposal ratio cancels, giving the "Metropolis" acceptance ratio $\min(1, \tilde{\pi}(\theta')/\tilde{\pi}(\theta^{(t)}))$.

Example — Theorem: Detailed balance and convergence of Metropolis–Hastings

The Metropolis–Hastings transition kernel $K(\theta, \theta') = q(\theta'|\theta)\alpha(\theta, \theta')$ (for $\theta' \neq \theta$, plus a "rejection mass" at $\theta' = \theta$) satisfies *detailed balance* with respect to π: $\pi(\theta)K(\theta, \theta') = \pi(\theta')K(\theta', \theta)$ for all $\theta \neq \theta'$. *Proof:* for $\theta' \neq \theta$, if $\frac{\tilde{\pi}(\theta')q(\theta|\theta')}{\tilde{\pi}(\theta)q(\theta'|\theta)} \leq 1$ (so $\alpha(\theta, \theta') =$ this ratio and $\alpha(\theta', \theta) = 1$), then $\pi(\theta)K(\theta, \theta') = \pi(\theta)q(\theta'|\theta)\frac{\tilde{\pi}(\theta')q(\theta|\theta')}{\tilde{\pi}(\theta)q(\theta'|\theta)} = \frac{\pi(\theta)}{\tilde{\pi}(\theta)}\tilde{\pi}(\theta')q(\theta|\theta')$, and since $\pi(\theta)/\tilde{\pi}(\theta) = \pi(\theta')/\tilde{\pi}(\theta')$ (both equal the reciprocal normalising constant), this equals $\pi(\theta')q(\theta|\theta') \times 1 = \pi(\theta')K(\theta', \theta)$, as required. Detailed balance implies π is a *stationary* distribution of the chain $(\int \pi(\theta)K(\theta, \theta')d\theta = \pi(\theta')$, by integrating detailed balance over θ). Combined with irreducibility and aperiodicity of the resulting chain (ensured by suitable choice of q, e.g. a Gaussian random walk with full support), the ergodic theorem for Markov chains gives $\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)}) \to \mathbb{E}_\pi[g(\theta)]$ almost surely, for any π-integrable g – the theoretical justification for using MCMC samples to estimate posterior expectations (Algorithm 20.1, step 2).

Hamiltonian Monte Carlo. HMC augments θ with auxiliary "momentum" variables $\mathbf{p} \sim \mathcal{N}(0, \mathbf{M})$ (mass matrix $\mathbf{M}$), defines a "Hamiltonian" $H(\theta, \mathbf{p}) = -\log\tilde{\pi}(\theta) + \frac{1}{2}\mathbf{p}^\top\mathbf{M}^{-1}\mathbf{p}$, and proposes new states by simulating Hamiltonian dynamics $\dot{\theta} = \mathbf{M}^{-1}\mathbf{p}, \dot{\mathbf{p}} = \nabla\log\tilde{\pi}(\theta)$ for some time (via a symplectic numerical integrator, typically the "leapfrog" scheme) before

applying a Metropolis accept/reject step based on H's change (which is exactly zero for exact Hamiltonian dynamics, and small but non-zero for the leapfrog discretisation, with the accept/reject step correcting for this discretisation error exactly). The key practical advantage: by following the gradient $\nabla \log \tilde{\pi}$, HMC proposals can move large distances in θ-space while maintaining high acceptance probability, dramatically improving mixing for high-dimensional, correlated posteriors – such as the $(\xi_j, \sigma_j, \mu_\xi, \tau_\xi^2, \dots)$ posteriors of the hierarchical model (Algorithm 20.3), where naive random-walk Metropolis (Algorithm 20.1) can mix very slowly due to the strong posterior correlations between ξ_j and the hyperparameters (μ_ξ, τ_ξ^2) induced by the shrinkage structure (Section 20.3's theorem).

Example — Worked example: random-walk Metropolis for a single GPD parameter

To sample $\xi \,|\, \text{data}$ for a GPD with known $\sigma = 2$ and $n = 50$ exceedances $\{e_i\}$, prior $\xi \sim \mathcal{N}(0.3, 0.15^2)$ (Section 20.1's worked example): at each step, propose $\xi' = \xi^{(t)} + \mathcal{N}(0, 0.05^2)$ (a random-walk step with standard deviation 0.05, a typical "tuned" step size for this kind of posterior). Compute $\log \tilde{\pi}(\xi') - \log \tilde{\pi}(\xi^{(t)}) = \left[\sum_i \log f_{\text{GPD}}(e_i; \xi', 2) + \log \phi_{0.3,0.15}(\xi')\right] - \left[\sum_i \log f_{\text{GPD}}(e_i; \xi^{(t)}, 2) + \log \phi_{0.3,0.15}(\xi^{(t)})\right]$ (log-likelihood plus log-prior, the GPD density $f_{\text{GPD}}(e; \xi, \sigma) = \sigma^{-1}(1 + \xi e/\sigma)^{-1/\xi - 1}$). Accept with probability $\min(1, e^{\log \tilde{\pi}(\xi') - \log \tilde{\pi}(\xi^{(t)})})$. A step size of 0.05 (roughly half the prior standard deviation 0.15, and typically tuned to achieve an acceptance rate of 20–40% for random-walk Metropolis in 1D) would, after a burn-in of a few hundred iterations, produce a chain whose empirical distribution approximates $\pi(\xi \,|\, \text{data})$ – the posterior $\mathbb{E}[\xi|\text{data}] \approx 0.32$, sd ≈ 0.10 quoted in Section 20.1's worked example would be the sample mean and standard deviation of this chain's (post-burn-in) values.

Convergence diagnostics. Two standard diagnostics, essential for any application of Algorithms 20.1, 20.3, 20.4: (1) the *potential scale reduction factor* $\hat{R}$ (Gelman–Rubin statistic), computed by running multiple (≥ 4) independent chains from over-dispersed starting points and comparing within-chain to between-chain variance – $\hat{R} \approx 1$ (conventionally $\hat{R} < 1.01$ or < 1.05) indicates the chains have converged to the same distribution; (2) *effective sample size* (ESS, conceptually analogous to but distinct from the importance-sampling ESS of Algorithm 19.1), accounting for autocorrelation within each chain – a chain of $T = 10000$ iterations with high autocorrelation might have ESS ≈ 200, meaning the posterior summary statistics have the precision of only 200 independent draws, not 10000. Both diagnostics should be checked routinely; failing either is a signal to run longer chains, re-tune the proposal (or switch to HMC), or re-parametrise the model (e.g. the "non-centred parametrisation" of hierarchical models, which re-expresses $\xi_j = \mu_\xi + \tau_\xi \tilde{\xi}_j$ for $\tilde{\xi}_j \sim \mathcal{N}(0, 1)$, often dramatically improving mixing for hierarchical models like Algorithm 20.3's, by reducing the strong posterior correlation between ξ_j and (μ_ξ, τ_ξ) that arises in the "centred" parametrisation, especially when τ_ξ is small).

Appendix F Reference Tables: GPD, GEV, and Related Distributions

Overview

This appendix collects, in one place, the density, distribution, and quantile functions for the Generalized Pareto Distribution (GPD) and related distributions used throughout Parts 19–20, together with their moments (where they exist) and the precise conditions for moment existence – supplementing Volume I's more extensive treatment, which this volume's Part 20 builds upon.

Generalized Pareto Distribution (GPD)

Density: $f(x;\xi,\sigma) = \frac{1}{\sigma}\left(1+\frac{\xi x}{\sigma}\right)^{-1/\xi-1}$ for $x \geq 0$ (if $\xi \geq 0$) or $0 \leq x \leq -\sigma/\xi$ (if $\xi < 0$), and $f(x;0,\sigma) = \frac{1}{\sigma}e^{-x/\sigma}$ (the $\xi \to 0$ limit, recovering the Exponential distribution).
CDF: $F(x;\xi,\sigma) = 1-(1+\xi x/\sigma)^{-1/\xi}$ ($\xi \neq 0$), $F(x;0,\sigma) = 1-e^{-x/\sigma}$.
Quantile function: $x_q = \frac{\sigma}{\xi}[(1-q)^{-\xi}-1]$ ($\xi \neq 0$), $x_q = -\sigma\log(1-q)$ ($\xi = 0$).
Mean: $\mathbb{E}[X] = \frac{\sigma}{1-\xi}$ for $\xi < 1$ (infinite or undefined for $\xi \geq 1$).
Variance: $\mathrm{Var}(X) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}$ for $\xi < 1/2$ (infinite or undefined for $\xi \geq 1/2$).
k-th moment: $\mathbb{E}[X^k] < \infty$ iff $\xi < 1/k$ – the GPD has only finitely many finite moments for $\xi > 0$, and *no* finite moments at all for $\xi \geq 1$ (not even the mean).

Generalized Extreme Value (GEV) Distribution

CDF: $F(x;\xi,\mu,\sigma) = \exp\left(-\left[1+\xi\frac{x-\mu}{\sigma}\right]^{-1/\xi}\right)$ for $1+\xi(x-\mu)/\sigma > 0$ ($\xi \neq 0$); $F(x;0,\mu,\sigma) = \exp(-e^{-(x-\mu)/\sigma})$ (Gumbel, $\xi = 0$ limit).
Quantile function: $x_q = \mu + \frac{\sigma}{\xi}\left[(-\log q)^{-\xi}-1\right]$ ($\xi \neq 0$); $x_q = \mu - \sigma\log(-\log q)$ ($\xi = 0$).
Relation to GPD: if block maxima follow GEV(ξ,μ,σ), threshold exceedances (above a sufficiently high threshold) follow approximately GPD(ξ,σ_u) for some σ_u depending on the threshold – the *same* shape parameter ξ governs both the block-maxima (GEV) and threshold-exceedance (GPD) representations, a consequence of the Pickands–Balkema–de Haan theorem (the foundational result of Volume I's peaks-over-threshold methodology).

Example — Table: moment existence and tail classification by ξ

- $\xi < 0$: **bounded tail** (Weibull-type GEV / Beta-type GPD). Support is bounded above, $[0, -\sigma/\xi]$ for the GPD. All moments exist. Example: $\xi = -0.2$ describes, e.g., certain wind-speed or wave-height maxima with a physical upper bound.
- $\xi = 0$: **exponential tail** (Gumbel-type GEV / Exponential GPD). All moments exist ($\mathbb{E}[X^k] = k!\sigma^k$). Example: many "moderately heavy" insurance loss distributions are well-approximated by $\xi \approx 0$.
- $0 < \xi < 1/2$: **heavy tail, finite mean and variance** (Fréchet-type GEV / Pareto-type GPD). Example: $\xi \approx 0.2$–0.3 is typical for operational risk loss severities (Section 20.4's worked example, $\hat{\xi} \approx 0.35$, is at the upper end of this range).
- $\xi = 1/2$: variance becomes infinite; this is the boundary often associated with the "edge of the Gaussian central limit theorem domain of attraction" for sums.
- $1/2 \leq \xi < 1$: **finite mean, infinite variance**. Example: certain catastrophe loss distributions (e.g. some earthquake or pandemic loss models) have been estimated with ξ in this range, implying that sample variances of historical losses are not reliable estimates of any underlying "true" variance (which does not exist).
- $\xi \geq 1$: **infinite mean**. Extremely heavy tails; Section 20.1's remarks note that even a small posterior probability mass on $\xi \geq 1$ renders the posterior predictive *mean* infinite, though quantiles (for $q < 1$) remain finite and well-defined.

Relative entropy between GPDs (for prior/posterior comparison). For two GPDs with the same σ but different shapes ξ_1, ξ_2, the KL divergence $D(\mathrm{GPD}(\xi_1,\sigma)\|\mathrm{GPD}(\xi_2,\sigma))$ does not have a simple closed form in general, but for ξ_1, ξ_2 both close to a common value ξ_0, a second-

order Taylor expansion gives $D \approx \frac{1}{2}(\xi_1 - \xi_2)^2 \mathcal{J}(\xi_0)$ where $\mathcal{J}(\xi_0)$ is the Fisher information for ξ at σ fixed – for the GPD, $\mathcal{J}(\xi) = \frac{2}{(1+\xi)(1+2\xi)}$ (for $\xi > -1/2$), a formula relevant to Section 20.2's discussion of Jeffreys priors (which involve $\sqrt{\det \mathbf{I}}$) and to assessing, via the local relative-entropy approximation, how "different" two posterior ξ-estimates from different sources (Section 20.3) really are in information-theoretic terms – e.g. for $\xi_0 \approx 0.27$ (Section 20.3's worked example population mean), $\mathcal{J}(0.27) = \frac{2}{1.27 \times 1.54} \approx 1.023$, so two sources with $\hat{\xi}_1 = 0.45, \hat{\xi}_2 = -0.05$ (a difference of 0.5) correspond to $D \approx \frac{1}{2}(0.5)^2 \times 1.023 \approx 0.128$ nats – a "moderate" information-theoretic distance, smaller than, e.g., the Section 17.3 exercise's loaded-die divergence of 0.511 nats, providing a rough calibration for "how different" these two sources' tail behaviours really are.

Bibliography

References

Foundational large deviation theory

Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles*, 736, 5–23.

Sanov, I. N. (1957). On the probability of large deviations of random variables. *Matematicheskii Sbornik*, 42, 11–44.

Varadhan, S. R. S. (1966). Asymptotic probabilities and differential equations. *Communications on Pure and Applied Mathematics*, 19(3), 261–286.

Gärtner, J. (1977). On large deviations from the invariant measure. *Theory of Probability and Its Applications*, 22(1), 24–39.

Ellis, R. S. (1984). Large deviations for a general class of random vectors. *Annals of Probability*, 12(1), 1–12.

Dembo, A. & Zeitouni, O. (1998). *Large Deviations Techniques and Applications* (2nd ed.). Springer.

Deuschel, J.-D. & Stroock, D. W. (1989). *Large Deviations*. Academic Press.

Freidlin–Wentzell theory and metastability

Freidlin, M. I. & Wentzell, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer.

Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4), 284–304.

Eyring, H. (1935). The activated complex in chemical reactions. *Journal of Chemical Physics*, 3(2), 107–115.

Bovier, A. & den Hollander, F. (2015). *Metastability: A Potential-Theoretic Approach*. Springer.

Donsker, M. D. & Varadhan, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I–II. *Communications on Pure and Applied Mathematics*, 28(1), 1–47; 28(2), 279–301.

Sethian, J. A. & Vladimirsky, A. (2003). Ordered upwind methods for static Hamilton–Jacobi equations: theory and algorithms. *SIAM Journal on Numerical Analysis*, 41(1), 325–363.

Scheffer, M., Bascompte, J., Brock, W. A., et al. (2009). Early-warning signals for critical transitions. *Nature*, 461, 53–59.

Kuehn, C. (2011). A mathematical framework for critical transitions: bifurcations, fast–slow time scales and stochastic dynamics. *Physica D*, 240(12), 1020–1035.

Rare-event simulation

Asmussen, S. & Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer.
Bucklew, J. A. (2004). *Introduction to Rare Event Simulation*. Springer.
Au, S.-K. & Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4), 263–277.
Rubinstein, R. Y. (1997). Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1), 89–112.
Rubinstein, R. Y. & Kroese, D. P. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer.
Cérou, F. & Guyader, A. (2007). Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2), 417–443.
Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3), 411–436.
Rubino, G. & Tuffin, B. (Eds.) (2009). *Rare Event Simulation using Monte Carlo Methods*. Wiley.
Kahn, H. & Harris, T. E. (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12, 27–30.

Bayesian inference and computation
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.
Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
O'Hagan, A., Buck, C. E., Daneshkhah, A., et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley.
James, W. & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.
Efron, B. & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5), 119–127.
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4), 382–401.
Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, 13(3), 917–1007.

Extreme value theory and catastrophe modelling
Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.
Coles, S. G. & Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Applied Statistics*, 45(4), 463–478.
Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.
Mitchell-Wallace, K., Jones, M., Hillier, J., & Foote, M. (Eds.) (2017). *Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide*. Wiley Blackwell.

Stochastic calculus, control, and convex analysis
Fleming, W. H. & Soner, H. M. (2006). *Controlled Markov Processes and Viscosity Solutions* (2nd ed.). Springer.
Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.
Seneta, E. (2006). *Non-negative Matrices and Markov Chains* (2nd ed.). Springer.
Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.

Appendix G Extended Case Study: An End-to-End Worked Portfolio

Overview

This appendix works a single numerical scenario through every Part of the volume, in sequence, to show how the separate pieces of machinery – rate functions (Part 17), metastability (Part 18), rare-event simulation (Part 19), and Bayesian tail inference (Part 20) – combine on one concrete problem: a proprietary trading desk's annual P&L, modelled first as an i.i.d. sum, then as a regime-switching process with a metastable "drawdown" state, then analysed computationally for its extreme quantiles, and finally re-estimated under parameter uncertainty.

The scenario

A trading desk's daily P&L (in \$M) is modelled, in a first approximation, as i.i.d. $X_i \sim \mathcal{N}(0.02, 0.30^2)$ (mean \$20K, sd \$300K per day). Over $n = 250$ trading days, the annual P&L is $S_{250} = \sum_{i=1}^{250} X_i$, with $\mathbb{E}[S_{250}] = 5.0$ (\$5M expected annual profit) and $\mathrm{sd}(S_{250}) = 0.30\sqrt{250} \approx 4.74$ (\$4.74M annual sd). The desk's risk committee asks four questions, addressed in turn below using Parts 17–20's tools.

Example — Question 1 (Part 17): probability of an annual loss exceeding \$5M

Using Cramér's theorem (Section 17.2) for $\bar{X}_{250} = S_{250}/250$, with $\Lambda(\lambda) = 0.02\lambda + \frac{1}{2}(0.30)^2\lambda^2$ (Gaussian CGF), $\Lambda^\star(x) = \frac{(x-0.02)^2}{2\times 0.09} = \frac{(x-0.02)^2}{0.18}$. The target $S_{250} \le -5$ corresponds to $\bar{X}_{250} \le -0.02$: $\Lambda^\star(-0.02) = \frac{(-0.02-0.02)^2}{0.18} = \frac{0.0016}{0.18} \approx 0.00889$. So $\Pr(\bar{X}_{250} \le -0.02) \approx e^{-250\times 0.00889} = e^{-2.22} \approx 0.108$ – an $\sim 11\%$ annual probability of losing $\ge$ \$5M, a sobering but not extreme figure for a desk running at this risk/return ratio (Sharpe ratio $\approx 0.02/0.30 \approx 0.067$ daily, ≈ 1.05 annualised). For comparison, the exact Gaussian probability is $\Pr(\mathcal{N}(5, 4.74^2) \le -5) = \Pr(Z \le -10/4.74) = \Pr(Z \le -2.11) \approx 0.0174$ – noticeably smaller than the LDP approximation 0.108. This discrepancy is expected: the LDP rate $e^{-nI(x)}$ captures the *exponential* rate correctly, but for $n = 250$ (not asymptotically large) the polynomial prefactor (here, related to the Gaussian normalisation $1/\sqrt{2\pi n\sigma^2}$) is far from negligible – exactly the "prefactor matters at finite n" caveat flagged in Section 18.4's discussion of Eyring–Kramers, here in its simplest (Gaussian, $n = 250$) incarnation. The exact answer (0.0174) should be preferred when available; the LDP estimate's value lies in situations where no exact formula exists.

Example — Question 2 (Part 18): regime-switching extension and drawdown persistence

The risk committee notes that the i.i.d. model ignores "drawdown persistence" – bad days tend to cluster. A refined model: a hidden 2-state chain (state 0 ="normal," state 1 ="stressed," transition matrix $\mathbf{P} = \begin{pmatrix} 0.98 & 0.02 \\ 0.10 & 0.90 \end{pmatrix}$, so stressed periods last $\sim 1/0.10 = 10$ days on average and occur roughly every $1/0.02 = 50$ days), with $X_i|\text{state} = 0 \sim \mathcal{N}(0.03, 0.25^2)$ and $X_i|\text{state} = 1 \sim \mathcal{N}(-0.10, 0.50^2)$. The stationary distribution solves $\pi\mathbf{P} = \pi$: $\pi_0 \times 0.02 = \pi_1 \times 0.10$, so $\pi_1/\pi_0 = 0.2$, giving $\pi \approx (0.833, 0.167)$ – the desk spends $\approx 17\%$ of days "stressed." The long-run mean daily P&L is $0.833 \times 0.03 + 0.167 \times (-0.10) \approx 0.025 - 0.0167 = 0.0083$ – noticeably lower than the i.i.d. model's 0.02, illustrating that ignoring regime-switching can *over-*

state the mean even when each regime's mean is individually plausible.

Quasi-potential interpretation (Section 18.2). Treating "stressed $\to$ normal" and "normal $\to$ stressed" as analogous to exits from two metastable wells (with "well depths" related to $-\log(\text{exit probability per day})$): the normal-state well depth is $-\log(0.02) \approx 3.91$ and the stressed-state well depth is $-\log(0.10) \approx 2.30$ – the stressed state is "shallower" (easier to exit), consistent with its shorter average duration. A risk manager wanting *shorter* stressed periods wants a *shallower* stressed-state well (higher exit rate), while one wanting *rarer* entries into the stressed state wants a *deeper* normal-state well (lower entry rate 0.02) – two distinct, independently-controllable design goals that the quasi-potential language helps keep conceptually separate.

Example — Question 3 (Part 19): simulating the 1-in-100-year annual loss

The committee wants the 99% annual loss quantile under the regime-switching model – a "1-in-100-year" figure. Direct Monte Carlo: simulate N years, each a 250-day regime-switching path, compute S_{250} for each, take the 1% empirical quantile. For $N = 10^4$ years, the 1% quantile is determined by the ~ 100 worst years – a reasonable but not generous effective sample size. If instead the committee wants the 99.9% quantile, $N = 10^4$ gives only ~ 10 years in the relevant tail – insufficient for a stable estimate, motivating either $N = 10^6$ or an importance-sampling/AMS approach (Sections 19.2–19.5).

AMS sketch. Using $\Phi(\text{year}) = -S_{250}$ as the score function, AMS (Algorithm 19.5) with $N = 500$ particles targets the 99.9% quantile ($p = 0.001$, $|\log p| \approx 6.9$) with RE $\approx \sqrt{6.9/500} \approx 0.117$ – an $\sim 12\%$ relative error from just 500 particles' worth of simulation, versus the $\sim 10^6$ direct-simulation years that would be needed for comparable precision. The "branch point" for this path-space AMS application would be the day $t^\star$ at which a discarded particle's running quantity $-S_t$ first reached the current level a – branching means: take a surviving particle's path up to the *same calendar day* $t^\star$, a subtlety in path-space AMS requiring care, often resolved by defining Φ via a first-passage-time formulation.

Example — Question 4 (Part 20): Bayesian re-estimation given 3 years of live data

After 3 years (750 trading days) of live trading, the desk has realised regime-classification data: 128 days classified as "stressed," giving empirical $\hat{\pi}_1 \approx 0.171$ – close to but not identical to the model's $\pi_1 \approx 0.167$. A Bayesian update of $(\alpha, \beta) = (0.02, 0.10)$ given the observed state sequence would use a Beta-Binomial conjugate update: if the data show 12 "normal$\to$stressed" transitions out of 622 normal-state days (empirical $\hat{\alpha} \approx 0.0193$, close to 0.02) and 58 "stressed$\to$normal" transitions out of 128 stressed-state days (empirical $\hat{\beta} \approx 0.453$, *far* from 0.10!) – this large discrepancy in $\hat{\beta}$ (implying stressed periods last ~ 2.2 days, not ~ 10) would, via the hierarchical/Bayesian updating of Section 20.3's machinery (here in its Beta-Binomial form), produce a posterior for β shifted substantially toward 0.453, with the shift's magnitude depending on the prior's strength. The practical upshot: 3 years of data suggest the desk's stressed periods are considerably *shorter* than the model assumed – a material recalibration that would flow through to a revised quasi-potential analysis (Question 2) and a revised 99.9% quantile (Question 3), illustrating the full Part-17-through-20 pipeline operating as a single, iteratively-refined risk assessment in which each Part's output informs the next Part's inputs.

Appendix H Notation and Symbol Glossary

Overview

This glossary collects the principal symbols used across Parts 17–20, with a brief description and a pointer to the section in which each is introduced. Symbols with highly local meaning are generally omitted; symbols reused across multiple Parts with a consistent meaning are prioritised.

Example — General probability and asymptotics

- $\Pr(\cdot)$, $\mathbb{E}[\cdot]$, $\mathrm{Var}(\cdot)$ – probability, expectation, variance under the relevant (context-specified) measure.
- ν, ρ, μ – probability measures/distributions (Section 17.1 onward); ν typically the "reference" or "true" law, ρ a candidate/empirical law.
- a_n – the LDP "speed" sequence, almost always n (Parts 17, 18.3) or $1/\epsilon$ (Part 18.1–18.2, 18.4–18.5).
- $I(x)$, $I_{\mathcal{X}}(x)$ – rate function (Section 17.1); subscripts indicate the space when multiple rate functions are in play (Section 17.5, contraction principle).
- $\approx$ – "agreement to leading exponential order," i.e. $\frac{1}{a_n}\log f \to \frac{1}{a_n}\log g$ (Section 17.1).

Example — Part 17 specific: CGFs, entropy, Markov chains

- $\Lambda(\lambda)$ – cumulant generating function (CGF), $\log \mathbb{E}[e^{\lambda X}]$ (Section 17.2).
- $\Lambda^\star(x)$ – Legendre–Fenchel transform of Λ (Section 17.2; Appendix A for full properties).
- $\mathcal{D}_\Lambda$ – domain of Λ, $\{\lambda : \Lambda(\lambda) < \infty\}$ (Appendix A).
- $\lambda^\star, \lambda^\star(x)$ – the optimal/LDP tilt parameter, solving $\Lambda'(\lambda^\star) = x$ (Sections 17.2, 19.2).
- ν_λ – exponentially tilted measure, $\nu_\lambda(dx) \propto e^{\lambda x}\nu(dx)$ (Section 19.2).
- $D(\rho\|\nu)$ – Kullback–Leibler divergence / relative entropy, $\sum_j \rho_j \log(\rho_j/\nu_j)$ (Section 17.3).
- $H(\rho)$ – Shannon entropy, $-\sum_j \rho_j \log \rho_j$ (Section 17.3, Appendix C).
- $\hat{\nu}_n$ – empirical measure of n i.i.d. samples (Section 17.3).
- Δ_K – probability simplex on K outcomes (Section 17.3, Appendix C).
- $\mathbf{P}$, $\mathbf{P}_\lambda$ – Markov transition matrix and its exponential tilt (Section 17.4; Appendix B for Perron–Frobenius theory).
- $\rho(\cdot)$ – spectral radius / Perron eigenvalue (Section 17.4, Appendix B) – *not* to be confused with ρ as a probability vector (Section 17.3) or correlation (elsewhere); context disambiguates.
- f – a continuous map in the contraction principle (Section 17.5), or a "reward"/observable function on a Markov chain's state space (Sections 17.4, 18.3).

Example — Part 18 specific: diffusions, action functionals, quasi-potentials

- X_t^ϵ – small-noise diffusion, $dX_t^\epsilon = b(X_t^\epsilon)dt + \sqrt{\epsilon}\sigma(X_t^\epsilon)dW_t$ (Section 18.1).
- $b, \sigma, a = \sigma\sigma^\top$ – drift, diffusion coefficient, and diffusion matrix (Section 18.1).

- ϵ – small-noise parameter (Sections 18.1–18.5).
- $I_{[0,T]}(\phi)$ – Freidlin–Wentzell action functional for path ϕ over $[0,T]$ (Section 18.1).
- $\phi, \phi^\star$ – a path / the minimal-action path (Section 18.1).
- $V(x)$, $V^\star$ – quasi-potential, and its minimum over a domain boundary ∂D (Section 18.2).
- $x^\star, x^\dagger$ – a stable equilibrium (well) and a saddle point (Sections 18.2, 18.4).
- τ^ϵ – first exit time from a domain (Section 18.2).
- $\mathbf{H}_\star, \mathbf{H}_\dagger$ – Hessians of the potential U at $x^\star, x^\dagger$ (Section 18.4).
- $U(x)$ – potential function, $b = -\nabla U$ in the gradient case (Section 18.2).
- $\mu_\dagger$ – magnitude of the (unique) negative eigenvalue of $\mathbf{H}_\dagger$ (Section 18.4).
- s, s_c – a slow control parameter and its critical (bifurcation) value (Section 18.5).
- $H(\theta, \mathbf{p})$ – (Appendix E only) the HMC Hamiltonian, $-\log\tilde{\pi}(\theta) + \frac{1}{2}\mathbf{p}^\top\mathbf{M}^{-1}\mathbf{p}$ – unrelated to the entropy $H(\rho)$ of Part 17 or the specific relative entropy $H(\mu|\mu P)$ of Section 18.3; context disambiguates the three uses of "H".

Example — Part 19 specific: importance sampling and simulation

- $q, q^\star, q_\theta$ – proposal distribution; zero-variance proposal; parametric proposal family (Sections 19.1, 19.4).
- $L(x)$ – likelihood ratio, $\nu(x)/q(x)$ (Section 19.1).
- $\hat{p}, \hat{p}_N$ – (importance-sampling or simulation) estimator of a probability p (Sections 19.1–19.5).
- RE – relative error, $\sqrt{\mathrm{Var}(\hat{p})}/(p\sqrt{N})$ or analogous (Section 19.1).
- ESS – effective sample size; distinct meanings in Algorithm 19.1 (importance-sampling ESS, $(\sum L_i)^2/\sum L_i^2$) vs. Appendix E (MCMC ESS, accounting for autocorrelation) – both measure "how many independent samples' worth of information," but via different mechanisms.
- $A, A_k, A^{(t)}$ – target rare event; intermediate level sets in subset simulation (Section 19.3) or cross-entropy (Section 19.4).
- p_0, ρ (Section 19.4 context) – level probability / elite fraction; *not* the same ρ as Part 17's probability vectors.
- γ_t, a_k – adaptive level thresholds (Sections 19.3, 19.4).
- Φ – score function (Sections 19.3, 19.5).
- N – generic sample size / particle count, with meaning specific to context.

Example — Part 20 specific: Bayesian inference and GPD/GEV

- ξ, σ – GPD shape and scale parameters (Appendix F for full distributional reference).
- $\pi(\theta), \pi(\theta|\text{data})$ – prior and posterior densities for parameter(s) θ (generic Bayesian notation, Sections 20.1–20.4, Appendix E).
- x_q – the q-quantile of a distribution (Appendix F).
- μ_ξ, τ_ξ^2 (and analogously $\mu_\sigma, \tau_\sigma^2$) – hyperparameters of the hierarchical model's population distribution for ξ_j (Section 20.3).
- v_j – sampling variance of source j's point estimate $\hat{\xi}_j$ (Section 20.3).

- M_k, $\Pr(M_k|\text{data})$ – candidate model k and its posterior probability (Section 20.4).
- Δ – generic "quantity of interest" in BMA (e.g. a return level) (Section 20.4).
- J – number of pooled sources/regions (Section 20.3) or candidate models (Section 20.4, where K is more common); Section 20.5 uses $J = 4$ for the number of catastrophe-model regions.
- $L_{j,i}, \Lambda_{\text{agg}}$ – per-event regional loss and aggregate annual loss (Section 20.5) – note Λ_{agg} is unrelated to the CGF Λ of Part 17 beyond the shared letter; the volume's organisation prevents ambiguity in context.

Appendix I Advanced Exercises

Overview

The exercises below extend Sections 17.1–20.5's per-section exercises with ten cross-cutting problems, each drawing on material from at least two Parts of the volume. Full solutions are provided; readers are encouraged to attempt each problem before reading the solution.

Example — Problem I.1 (Parts 17 & 19): tilting for a Poisson rare event

For $X \sim \text{Poisson}(\mu = 5)$, use exponential tilting (Sections 17.2, 19.2) to design an importance-sampling scheme for $\Pr(\bar{X}_{20} \geq 10)$ (i.e. $S_{20} \geq 200$, double the mean of 100).

Solution. From Section 17.2's exercises, $\Lambda^\star(x) = x\log(x/\mu) - x + \mu$ for Poisson(μ), with optimal tilt $\lambda^\star = \log(x/\mu)$. For $x = 10, \mu = 5$: $\lambda^\star = \log 2 \approx 0.693$, $\Lambda(\lambda^\star) = \mu(e^{\lambda^\star} - 1) = 5(2-1) = 5$. The tilted distribution $\nu_{\lambda^\star}(k) = e^{\lambda^\star k - 5} \times \frac{5^k e^{-5}}{k!} = \frac{(5e^{0.693})^k e^{-5} e^{-5}}{k!} = \frac{10^k e^{-10}}{k!}$ – exactly Poisson(10), confirming Poisson is closed under exponential tilting with $\mu \to \mu e^{\lambda^\star}$, analogous to the Gamma case of Section 19.2's exercises. **Algorithm:** simulate $X_1, \ldots, X_{20} \overset{\text{iid}}{\sim} \text{Poisson}(10)$ (mean 10, so $\bar{X}_{20} \approx 10$ is now typical), weight each replicate by $L = e^{-\lambda^\star S_{20} + 20\Lambda(\lambda^\star)} = e^{-0.693 \times S_{20} + 100}$. For a typical tilted sample with $S_{20} \approx 200$, $L \approx e^{-138.6+100} = e^{-38.6}$, and $\Pr(\bar{X}_{20} \geq 10) \approx e^{-20\Lambda^\star(10)} = e^{-20 \times (10\log 2 - 5)} = e^{-20 \times 1.931} = e^{-38.6}$ – matching, as expected.

Example — Problem I.2 (Parts 17 & 20): combining CGFs of two desks

Two trading desks have daily P&L CGFs $\Lambda_1(\lambda) = 0.01\lambda + \frac{1}{2}(0.2)^2\lambda^2$ and $\Lambda_2(\lambda) = 0.03\lambda + \frac{1}{2}(0.4)^2\lambda^2$ (both Gaussian). If a risk manager models the *combined* desk (sum of the two P&Ls, independent), what is $\Lambda_{\text{combined}}(\lambda)$, and how does $\Lambda^\star_{\text{combined}}$ compare to $\Lambda_1^\star, \Lambda_2^\star$?

Solution. For independent sums, CGFs add: $\Lambda_{\text{combined}}(\lambda) = \Lambda_1(\lambda) + \Lambda_2(\lambda) = 0.04\lambda + \frac{1}{2}(0.20)\lambda^2$ – $\mathcal{N}(0.04, 0.20)$ for the combined daily P&L (mean \$40K, sd $\approx$ \$447K). $\Lambda^\star_{\text{combined}}(x) = \frac{(x-0.04)^2}{0.40}$. Compare: $\Lambda_1^\star(x) = \frac{(x-0.01)^2}{0.08}$, $\Lambda_2^\star(x) = \frac{(x-0.03)^2}{0.32}$. At $x = 0$: $\Lambda_1^\star(0) \approx 0.00125$, $\Lambda_2^\star(0) \approx 0.00281$, $\Lambda^\star_{\text{combined}}(0) = 0.004$ – close to but not exactly $\Lambda_1^\star(0) + \Lambda_2^\star(0) \approx 0.00406$. In general, rate functions for sums of independent variables do *not* simply add, even though CGFs do – the rate function of a sum is the *infimal convolution* of the individual rate functions, $I_{1+2}(x) = \inf_{x_1+x_2=x}(I_1(x_1) + I_2(x_2))$, which for Gaussians reduces to the formula above via variance addition, but for non-Gaussian components requires genuinely solving the infimal convolution – a contraction-principle-

type calculation (Section 17.5) applied to the sum map $f(x_1, x_2) = x_1 + x_2$.

Example — Problem I.3 (Part 18): quasi-potential for an asymmetric well

For $U(x) = \frac{1}{4}x^4 - \frac{1}{3}x^3$, find the well, the saddle (if any), and $V(0)$.
Solution. $U'(x) = x^2(x-1) = 0$ at $x = 0, 1$. $U''(x) = 3x^2 - 2x = x(3x-2)$. $U''(1) = 1 > 0$: $x = 1$ is a local min (the well, $x^\star = 1$). $U''(0) = 0$: degenerate – $x = 0$ is neither a simple min nor a simple saddle in the Eyring–Kramers sense; Section 18.4's remarks on degenerate saddles apply. $U(1) = \frac{1}{4} - \frac{1}{3} = -\frac{1}{12}$, $U(0) = 0$. The quasi-potential $V(x) = 2[U(x) - U(1)]$ (gradient formula, Section 18.2): $V(0) = 2 \times \frac{1}{12} = \frac{1}{6}$. Since $x = 0$ is not a true (non-degenerate) saddle, the Eyring–Kramers *prefactor* formula does not directly apply there – a degenerate-saddle asymptotic (different power of ϵ in the prefactor) would be needed for a precise exit-time formula "through" $x = 0$, even though the quasi-potential *value* $V(0) = 1/6$ is well-defined via the gradient formula regardless of the degeneracy (the formula $V = 2\Delta U$ requires only a valid potential, not non-degenerate critical points – it is specifically the *prefactor* formula that requires non-degeneracy).

Example — Problem I.4 (Parts 19 & 20): CE method for a posterior tail probability

A posterior $\pi(\xi|\text{data}) \approx \mathcal{N}(0.3, 0.08^2)$. Estimate $\Pr(\xi > 0.5|\text{data})$ via a CE-based scheme (Section 19.4) and compare to the direct Gaussian-tail calculation.
Solution. Direct: $\Pr(\xi > 0.5) = \Pr(Z > (0.5 - 0.3)/0.08) = \Pr(Z > 2.5) \approx 0.0062$. CE scheme: family $q_\theta = \mathcal{N}(\theta, 0.08^2)$, target $A = \{\xi > 0.5\}$. After iterating (per Algorithm 19.4), $\theta^\star \to 0.5$ (the CE-optimal mean shift for a half-space event converges to the threshold, Section 19.4's exercises). With $q_{\theta^\star} = \mathcal{N}(0.5, 0.08^2)$, A has probability ≈ 0.5 under $q_{\theta^\star}$, and the importance weight for a "typical" tilted sample ($\xi \approx 0.5$) is $L = \frac{\phi_{0.3,0.08}(0.5)}{\phi_{0.5,0.08}(0.5)} = \frac{\phi(2.5)}{\phi(0)} = \frac{0.0175}{0.3989} \approx 0.0438$. A rough estimate $\hat{p} \approx 0.5 \times 0.0438 \approx 0.022$ is in the right ballpark of 0.0062 but not highly accurate from a single-weight approximation; a full IS run with $N' = 10^4$ samples from $q_{\theta^\star}$, each weighted by its own (sample-dependent) L, converges to the true 0.0062 – illustrating that even with a good tilt, the final estimate requires averaging over the weight distribution, not evaluating it at one point (cf. Algorithm 19.1's ESS diagnostic).

Example — Problem I.5 (Parts 17 & 18): from discrete to continuous large deviations

Explain the conceptual relationship between Cramér's theorem (discrete sums) and Schilder's theorem (continuous Brownian paths), and identify the "discrete analogue" of $I_{[0,T]}(\phi) = \frac{1}{2}\int_0^T |\dot{\phi}|^2 dt$.
Solution. Discretise $[0, T]$ into K steps of size $\Delta t = T/K$, $\phi_k = \phi(k\Delta t)$. The increments $\Delta\phi_k = \phi_k - \phi_{k-1}$ are, for $I_{[0,T]}(\phi) \approx \frac{1}{2}\sum_k \frac{(\Delta\phi_k)^2}{\Delta t}$, analogous to i.i.d. Gaussian increments $\mathcal{N}(0, \Delta t)$ under the reference measure – and $\frac{(\Delta\phi_k)^2}{\Delta t} = 2\Lambda^\star_{\mathcal{N}(0,\Delta t)}(\Delta\phi_k)$ (using $\Lambda^\star(x) = x^2/(2\sigma^2)$ for $\sigma^2 = \Delta t$). So $I_{[0,T]}(\phi) \approx \sum_k \Lambda^\star_{\mathcal{N}(0,\Delta t)}(\Delta\phi_k)$ – the path's action is the continuum limit of a sum of one-step Cramér rate functions, with each increment's "sample size" being 1 rather than $n \to \infty$. The action functional is the continuum (small-Δt) limit, with $K = T/\Delta t \to \infty$ playing the role $n \to \infty$ plays in Cramér's theorem – both arise from summing many independent small-variance Gaussian contributions, with the rate function in each case (twice) the Legendre transform of the

per-contribution CGF.

Example — Problem I.6 (Part 20): non-existence of the posterior mean

A posterior places probability 0.97 on $\mathcal{N}(0.3, 0.05^2)$ (truncated to $\xi < 1$) and 0.03 on Uniform$(1, 1.5)$. Is $\mathbb{E}[X_{n+1}|\text{data}]$ finite?
Solution. $\mathbb{E}[X_{n+1}|\text{data}] = \int \mathbb{E}[X|\xi,\sigma]\pi(\xi,\sigma|\text{data})d\xi\, d\sigma = \int \frac{\sigma}{1-\xi}\pi(\xi,\sigma|\text{data})d\xi\, d\sigma$ (GPD mean formula, Appendix F, valid for $\xi < 1$). For $\xi \geq 1$, $\mathbb{E}[X|\xi,\sigma]$ is genuinely *undefined* (the integral $\int x\, f(x;\xi,\sigma)dx$ diverges). By Fubini/Tonelli, $\mathbb{E}[X_{n+1}|\text{data}] = \mathbb{E}_{\xi,\sigma|\text{data}}[\mathbb{E}[X|\xi,\sigma]]$, and if $\mathbb{E}[X|\xi,\sigma] = +\infty$ on a set of positive posterior probability (here, 3%), the outer expectation is $+\infty$. So *any* positive posterior probability on $\xi \geq 1$ renders the posterior predictive mean infinite, regardless of how small that probability is – confirming Section 20.1's remarks, and motivating posterior predictive *quantiles* (always finite for $q < 1$, since $x_q(\xi,\sigma) = \frac{\sigma}{\xi}[(1-q)^{-\xi} - 1]$ is finite for all $\xi \neq 0$ including $\xi \geq 1$) as the primary risk summary whenever the posterior cannot rule out $\xi \geq 1$ with certainty.

Example — Problem I.7 (Parts 17, 19 & 20): a unified view of "rare"

The word "rare" appears in at least four senses: (a) a large-deviation event, $\Pr(\bar{X}_n \approx x) \approx e^{-nI(x)}$ (Part 17); (b) a noise-induced transition, $\mathbb{E}[\tau^\epsilon] \sim e^{V^\star/\epsilon}$ (Part 18); (c) a simulation target $p \to 0$ (Part 19); (d) an extreme quantile x_q, $q \to 1$ (Part 20). How do these relate?
Solution. All four are the same phenomenon – decay of probability as a parameter approaches an extreme – viewed through different lenses. In (a), $n \to \infty$ for fixed $x \neq \mathbb{E}[X]$. In (b), $\epsilon \to 0$ drives $\mathbb{E}[\tau^\epsilon] \to \infty$. In (c), p is a (small, fixed) *input* describing a sequence of problems of increasing difficulty, not itself a limiting probability. In (d), $q \to 1$ drives $x_q \to \infty$. The key qualitative distinction: (a) and (b) involve *exponential* rates ($I(x)$, $V^\star/\epsilon$), while (d) involves a *power-law* rate ($x_q \sim (1-q)^{-\xi}$ for $\xi > 0$) – reflecting that GPD/GEV tails are regularly varying rather than exponentially light, the heavy-tail regime explicitly excluded from Cramér/Gärtner–Ellis (Section 17.4's remarks). (c) is not a limiting statement but a computational response to whichever of (a), (b), (d) is operative. Recognising which sense is at play – sum large deviations, metastable transition, algorithm target, or tail quantile – is often the first step in selecting the right tool from this volume.

Example — Problem I.8 (Part 18): Girsanov verification of $I_{[0,T]}(\phi^\star) = 2\Delta U$

Verify, via Girsanov's theorem (Appendix D), that $I_{[0,T]}(\phi^\star) = 2\Delta U$ for the time-reversed gradient-ascent path $\dot{\phi}^\star = +U'(\phi^\star)$ (Section 18.1's worked example).
Solution. The control is $u(t) = \frac{1}{\sqrt{\epsilon}}(\dot{\phi}^\star - b(\phi^\star)) = \frac{1}{\sqrt{\epsilon}}(U'(\phi^\star) - (-U'(\phi^\star))) = \frac{2}{\sqrt{\epsilon}}U'(\phi^\star)$ (using $b = -U'$, $\dot{\phi}^\star = U'(\phi^\star)$). The Girsanov quadratic term is $\frac{1}{2}\int_0^T |u|^2 dt = \frac{2}{\epsilon}\int_0^T U'(\phi^\star)^2 dt$. Now $\int_0^T U'(\phi^\star)^2 dt = \int_0^T \dot{\phi}^\star \cdot U'(\phi^\star)dt = \int_0^T \frac{d}{dt}U(\phi^\star(t))dt = U(\phi^\star(T)) - U(\phi^\star(0)) = \Delta U$ (chain-rule identity, Section 18.1). So $\frac{1}{2}\int_0^T |u|^2 dt = \frac{2\Delta U}{\epsilon}$, giving $I_{[0,T]}(\phi^\star) = \epsilon \times \frac{2\Delta U}{\epsilon} = 2\Delta U$, matching Section 18.1 exactly – the ϵ's cancel because $I_{[0,T]}(\phi)$ is ϵ-independent, a property of $\phi^\star$ and U alone; ϵ enters only via $e^{-I/\epsilon}$.

Example — Problem I.9 (Parts 19 & 20): AMS applied to a posterior tail

Could AMS (Section 19.5) estimate $\Pr(\xi > 1|\text{data})$ (Problem I.6) when this probability is itself extremely small (e.g. 10^{-6}) and the posterior is known only via MCMC?
Solution. Yes, in principle. AMS requires (a) a score function $\Phi(\xi, \sigma) = \xi$, target $a_m = 1$, and (b) MCMC moves *conditional on* $\Phi \geq a$ for increasing a – i.e. sampling from $\pi(\xi, \sigma|\text{data}, \xi \geq a)$, e.g. via Metropolis-Hastings with rejection of proposals violating $\xi \geq a$. The key requirement – pointwise evaluability of the unnormalised posterior $\tilde{\pi}(\xi, \sigma) = L(\xi, \sigma|\text{data})\pi(\xi, \sigma)$ – is exactly what MCMC requires anyway, so AMS's branching step (regenerating a discarded particle from a branch point via constrained MCMC) can proceed using the same $\tilde{\pi}$. Since there is no "time"/path structure here (particles are just points (ξ, σ)), AMS degenerates to the static form of Algorithm 19.5's remarks – effectively subset simulation (Section 19.3) applied directly to the posterior, with $\Phi(\xi, \sigma) = \xi$. This is a genuinely useful "second-order" rare-event technique – rare events *about* a distribution describing rare events – arising naturally in the careful uncertainty quantification this volume advocates throughout Part 20.

Example — Problem I.10 (synthesis): designing a risk-monitoring dashboard

Sketch, using concepts from all four Parts, the key metrics a risk-monitoring dashboard might display, with each metric's source Part.
Solution. **(daily)** realised P&L vs. model-implied $\Lambda^\star$-based probability of that day's outcome (Part 17) – flags "surprising" days, a simple model-validation signal (cf. Section 17.3's χ^2-type validation note). **(weekly)** early-warning indicators – rolling variance and autocorrelation of key risk factors (Section 18.5) – flagging potential approach to a regime change before it occurs. **(monthly)** updated extreme-quantile estimates via analytic (Parts 17/20) and simulation-based (Part 19) methods, with explicit uncertainty bands from parameter uncertainty (Section 20.1) and model uncertainty (Section 20.4) – a number plus an uncertainty band, not a bare number. **(quarterly)** re-estimation of hierarchical/pooled tail parameters across business lines (Section 20.3), tracking whether each line's parameters move toward or away from the portfolio-wide population estimate. **(annual/trigger-based)** full re-assessment of dependence structures and BMA model weights (Sections 20.4–20.5), feeding back into the more-frequent metrics above. This layered structure – fast cheap checks (Part 17) at high frequency, slower judgement-intensive analyses (Parts 19–20) at low frequency, each informing the others – reflects a general principle for deploying this volume's mathematics operationally: not every tool runs at the same frequency, but all are part of one coherent system.

Appendix J Numerical Reference Tables and Topic Guide

Overview

This final appendix collects two kinds of reference material that readers may wish to consult repeatedly while working through Parts 17–20: numerical tables of rate functions and quantiles for the distributions used most often in this volume, and a topic-based guide indicating which sections, appendices, and bibliography entries are most relevant to a given practical question.

Example — Table J.1: Gaussian rate function $\Lambda^\star(x) = (x-\mu)^2/(2\sigma^2)$, standardised ($\mu = 0, \sigma = 1$)

- $x = 0.5$: $\Lambda^\star = 0.125$. $e^{-n\Lambda^\star}$ for $n = 10$: $e^{-1.25} \approx 0.287$; $n = 100$: $e^{-12.5} \approx 3.7 \times 10^{-6}$.
- $x = 1.0$: $\Lambda^\star = 0.5$. $n = 10$: $e^{-5} \approx 6.7 \times 10^{-3}$; $n = 100$: $e^{-50} \approx 2.0 \times 10^{-22}$.
- $x = 1.5$: $\Lambda^\star = 1.125$. $n = 10$: $e^{-11.25} \approx 1.3 \times 10^{-5}$; $n = 100$: $e^{-112.5} \approx 5 \times 10^{-49}$.
- $x = 2.0$: $\Lambda^\star = 2.0$. $n = 10$: $e^{-20} \approx 2.1 \times 10^{-9}$; $n = 100$: $e^{-200} \approx 10^{-87}$.
- $x = 3.0$: $\Lambda^\star = 4.5$. $n = 10$: $e^{-45} \approx 2.9 \times 10^{-20}$; $n = 100$: $e^{-450} \approx 10^{-196}$.

Reading the table. For non-standard $\mathcal{N}(\mu, \sigma^2)$, replace x by $(x-\mu)/\sigma$ (the standardised deviation) and read $\Lambda^\star$ from the table; the resulting $\Lambda^\star$ is then used directly in $e^{-n\Lambda^\star}$, since $\Lambda^\star_{\mathcal{N}(\mu,\sigma^2)}(x) = \Lambda^\star_{\mathcal{N}(0,1)}((x-\mu)/\sigma)$ for the Gaussian's scale-location family (a direct consequence of property (v) in Appendix A applied to the affine reparametrisation $x \mapsto (x-\mu)/\sigma$). These values were used throughout Section 17.2 and the Appendix G case study; e.g. Appendix G's $\Lambda^\star(-0.02) \approx 0.00889$ corresponds to a standardised deviation of $(-0.02 - 0.02)/0.30 \approx -0.133$, and indeed $\Lambda^\star_{\mathcal{N}(0,1)}(0.133) = (0.133)^2/2 \approx 0.00889$, consistent with the table's $x = 0.5$ row scaled down by $(0.133/0.5)^2 \approx 0.071$: $0.125 \times 0.071 \approx 0.00889$. ✓

Example — Table J.2: Exponential rate function $\Lambda^\star(x) = \theta x - 1 - \log(\theta x)$, standardised ($\theta = 1$, x in units of the mean)

- $x = 1.5$ ($1.5\times$ mean): $\Lambda^\star = 1.5 - 1 - \log 1.5 \approx 0.5 - 0.405 = 0.095$.
- $x = 2.0$ ($2\times$ mean): $\Lambda^\star = 2 - 1 - \log 2 \approx 1 - 0.693 = 0.307$.
- $x = 3.0$ ($3\times$ mean): $\Lambda^\star = 3 - 1 - \log 3 \approx 2 - 1.099 = 0.901$.
- $x = 5.0$ ($5\times$ mean): $\Lambda^\star = 5 - 1 - \log 5 \approx 4 - 1.609 = 2.391$.
- $x = 10.0$ ($10\times$ mean): $\Lambda^\star = 10 - 1 - \log 10 \approx 9 - 2.303 = 6.697$.

Reading the table. For $\mathrm{Exp}(\theta)$ with mean $1/\theta$, a target x in original units corresponds to θx in the table (i.e. "x as a multiple of the mean"). Note the qualitative contrast with Table J.1: the Gaussian $\Lambda^\star$ grows *quadratically* in the standardised deviation ($\Lambda^\star(2) = 2.0$, $\Lambda^\star(3) = 4.5$, ratio $2.25 = (3/2)^2$), while the Exponential $\Lambda^\star$ grows roughly *linearly* for large x ($\Lambda^\star(x) \approx x$ for large x, since $-1 - \log x$ becomes negligible relative to x) – so exponentially-tailed sums have *heavier* large-deviation tails than Gaussian sums of the same variance, at sufficiently extreme x – the point underlying Section 17.2's worked-example comparison at $x = 3$ (which used $\theta = 1$, matching this table directly: $\Lambda^\star(3) \approx 0.901$, as tabulated).

Example — Table J.3: GPD quantile ratios $x_{0.999}/x_{0.99}$ as a function of ξ (σ cancels in the ratio)

- $\xi = 0$ (Exponential): $x_q = -\log(1-q)$ (in units of σ). $x_{0.999}/x_{0.99} = \log(1000)/\log(100) = 6.908/4.605 \approx 1.50$.
- $\xi = 0.1$: $x_{0.999}/x_{0.99} = \frac{(0.001)^{-0.1}-1}{(0.01)^{-0.1}-1} = \frac{10^{0.3}-1}{10^{0.2}-1} = \frac{0.995}{0.585} \approx 1.70$.
- $\xi = 0.2$: $\frac{10^{0.6}-1}{10^{0.4}-1} = \frac{2.981}{1.512} \approx 1.97$.
- $\xi = 0.3$: $\frac{10^{0.9}-1}{10^{0.6}-1} = \frac{6.943}{2.981} \approx 2.33$.
- $\xi = 0.4$: $\frac{10^{1.2}-1}{10^{0.8}-1} = \frac{14.85}{5.31} \approx 2.80$.
- $\xi = 0.5$: $\frac{10^{1.5}-1}{10^{1.0}-1} = \frac{30.62}{9} \approx 3.40$.

Reading the table. This ratio – "how much does the loss escalate from a 1-in-100 to a 1-in-1000 event?" – is one of the most decision-relevant single numbers in extreme value risk modelling, and Table J.3 shows it is highly sensitive to ξ: a "$1.5\times$" escalation ($\xi = 0$) versus a "$3.4\times$" escalation ($\xi = 0.5$) for the *same* underlying $10\times$ change in return period. This table directly supports the elicitation discussion of Section 20.2: an expert's stated ratio between two return-period loss estimates *is*, via this table (read in reverse, finding ξ from the ratio), an implicit statement about ξ – though note Section 20.2's own worked example used the $(x_{0.9}, x_{0.99})$ pair rather than $(x_{0.99}, x_{0.999})$, so this table's specific numbers should not be applied to ratios computed from a different quantile pair without re-deriving the appropriate ratio formula by the same method.

Example — Table J.4: Eyring–Kramers exponential factor $e^{\Delta U/\epsilon}$ for various $(\Delta U, \epsilon)$

- $\Delta U = 0.25, \epsilon = 0.10$: $e^{2.5} \approx 12.2$.
- $\Delta U = 0.25, \epsilon = 0.05$: $e^{5} \approx 148.4$.
- $\Delta U = 0.25, \epsilon = 0.02$: $e^{12.5} \approx 2.68 \times 10^{5}$.
- $\Delta U = 0.50, \epsilon = 0.10$: $e^{5} \approx 148.4$.
- $\Delta U = 0.50, \epsilon = 0.05$: $e^{10} \approx 2.20 \times 10^{4}$.
- $\Delta U = 0.50, \epsilon = 0.02$: $e^{25} \approx 7.2 \times 10^{10}$.
- $\Delta U = 1.00, \epsilon = 0.05$: $e^{20} \approx 4.85 \times 10^{8}$.
- $\Delta U = 1.00, \epsilon = 0.02$: $e^{50} \approx 5.18 \times 10^{21}$.

Reading the table. The dramatic range spanned by this table (from ~ 12 to $\sim 10^{21}$) is the central quantitative fact underlying Sections 18.4–18.5 and Part 19's motivation: mean exit/escape/transition times vary by many orders of magnitude for what might seem like "modest" changes in barrier height ΔU or noise level ϵ – a $4\times$ change in ΔU (from 0.25 to 1.00, at fixed $\epsilon = 0.05$) changes the exponential factor from ~ 148 to $\sim 4.85 \times 10^{8}$, a factor of $\sim 3 \times 10^{6}$. This table also provides quick sanity-checks for Algorithm 18.4's numerical output: if a computed $(\Delta U, \epsilon)$ pair gives an exponential factor wildly inconsistent with interpolation from this table, a computational error (e.g. in identifying $x^{\dagger}$ or computing $U(x^{\dagger}) - U(x^{\star})$) is likely.

Topic guide: "I need to..."

...estimate how surprising an observed sample mean is, given a model. Sections 17.1–17.2 (rate function, Cramér's theorem), Algorithm 17.2 (numerical $\Lambda^{\star}$), Tables J.1/J.2 (quick lookups for Gaussian/Exponential), Appendix C (full proofs if rigour is required).
...assess whether an empirical distribution is consistent with an assumed model. Section 17.3 (Sanov's theorem), its application note (χ^2-test connection), Appendix C (full proof of Sanov's theorem).
...analyse a regime-switching or Markov-modulated process's long-run behaviour. Sections 17.4 (Gärtner–Ellis) and 18.3 (Donsker–Varadhan), Appendix B (Perron–Frobenius background), Algorithms 17.4/18.3 (numerical procedures), Appendix G (worked regime-switching case study).
...transfer a known rate function to a derived quantity (a function of the original variable). Section 17.5 (contraction principle), Algorithm 17.5.
...understand or compute the probability of a rare path/trajectory for a diffusion. Section 18.1 (Freidlin–Wentzell, action functional), Appendix D (Girsanov's theorem,

rigorous foundations), Algorithm 18.1.
...estimate a mean transition/exit/escape time for a metastable system. Sections 18.2 (quasi-potential) and 18.4 (Eyring–Kramers), Algorithms 18.2/18.4, Table J.4 (quick numerical lookups).
...detect early-warning signals of an approaching transition in time-series data. Section 18.5, Algorithm 18.5.
...estimate a small probability via simulation, when crude Monte Carlo is infeasible. Start with Section 19.1 (diagnosing the problem via RE and ESS), then Section 19.2 (if an analytic tilt is available, the cheapest option), else Section 19.3 (subset simulation, model-free) or Section 19.4 (cross-entropy, adaptive) or Section 19.5 (AMS, if a CLT-based confidence interval is needed). Algorithms 19.1–19.5 provide pseudocode for each.
...fit a GPD/GEV tail with proper uncertainty quantification. Section 20.1 (posterior predictive quantiles), Algorithm 20.1, Appendix E (MCMC background), Appendix F (distributional formulas).
...build a prior for a tail-shape parameter when data are scarce. Section 20.2, Algorithm 20.2.
...combine tail estimates across multiple related sources/regions/business lines. Section 20.3 (hierarchical pooling, shrinkage), Algorithm 20.3, Appendix G's Question 4 for a worked Beta-Binomial analogue.
...quantify the impact of not knowing which model is "correct." Section 20.4 (BMA, variance decomposition), Algorithm 20.4.
...build an integrated multi-region catastrophe/aggregate-loss model. Section 20.5, Algorithm 20.5, and (for dependence-structure mathematics specifically) the bibliography's "Extreme value theory and catastrophe modelling" section.
...understand the convex-analytic machinery underlying $\Lambda \leftrightarrow \Lambda^\star$. Appendix A.
...verify or extend a Markov-chain-based rate function calculation numerically. Appendix B, Algorithm 17.4/18.3.
...see a fully rigorous proof of Cramér's or Sanov's theorem. Appendix C.
...understand the stochastic-calculus foundations of Part 18. Appendix D.
...run or troubleshoot MCMC for any of Part 20's models. Appendix E.
...look up a GPD/GEV formula without searching back through Part 20. Appendix F, Table J.3.
...see how Parts 17–20 fit together on a single example. Appendix G.
...resolve a notational ambiguity (the same symbol used differently in different Parts). Appendix H.
...test understanding with problems spanning multiple Parts. Appendix I.

Closing remarks

Volume V has developed the mathematics of "how rare is rare, and what can be done about it" across four interlocking Parts: the rate-function calculus of Part 17 provides the basic vocabulary (exponential decay rates, governed by Legendre transforms and relative entropy); Part 18 extends this vocabulary to dynamic, path-dependent settings, where "rare" becomes "metastable" and the relevant question shifts from a single probability to a transition rate; Part 19 confronts the computational reality that rate functions, however elegant, often cannot be evaluated in closed form, and develops a suite of simulation techniques whose efficiency gains – from $O(1/p)$ to $O(|\log p|)$ – are among the most dramatic in all of computational statistics; and Part 20 closes the loop by asking how the parameters governing all of the above should themselves

be estimated, and with what honestly-reported uncertainty, from necessarily limited data.

The volume's appendices have aimed to make this material self-contained: Appendices A–D supply the convex-analytic, linear-algebraic, and stochastic-calculus prerequisites in full; Appendix E supplies the Bayesian-computational prerequisites; Appendix F is a quick-reference for the distributions used throughout; and Appendices G–J provide synthesis, notational clarity, additional practice, and numerical orientation. Together with Volumes I–IV and VI of the *Mathematics of Robustness, Risk & Failure* series, Volume V's large-deviation perspective provides one of the deepest – and, as the application notes throughout have aimed to show, one of the most practically consequential – mathematical lenses through which to understand the behaviour of complex systems under stress.